DOCUMENT RESUME

ED 395 005 TM 025 001

AUTHOR	Stocking, Martha L.; And Others
TITLE	Factors Affecting the Sample Invariant Properties of Linear and Curvilinear Observed- and True-Score Equating Procedures.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RR-88-41
PUB DATE	Aug 88
NOTE	95p.; Version of a paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 1988).
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC04 Plus Postage.
DESCRIPTORS	*Equated Scores; *Item Response Theory; *Sample Size; Simulation; *True Scores
IDENTIFIERS	Curvilinear Functions; Equipercentile Equating; Invariance; Levine Equating Method; Tucker Common Item Equating Method

ABSTRACT

A sequence of simulations was carried out to aid in the diagnosis and interpretation of equating differences found between random and matched (nonrandom) samples for four commonly used equating procedures: (1) Tucker linear observed-score equating; (2) Levine equally reliable linear observed-score equating; (3) equipercentile curvilinear observed-score equating; and (4) item response theory (IRT) curvilinear true-score equating. The results support the prediction based on theoretical grounds that observed-score equating methods are more affected by sample variation than are true-score equating methods. These results further suggest that matching equating samples on the basis of fallible measures of ability may not be advisable for any conventional equating method except the Tucker method. In addition, the results support a particular hypothesis about IRT equating, suggesting that the use of matched samples cannot be recommended for this equating method either. (Contains 8 tables, 8 figures, and 12 references.) (Author/SLD)

RR-88-41

RESEARCH REPORT

FACTORS AFFECTING THE SAMPLE INVARIANT PROPERTIES OF LINEAR AND CURVILINEAR OBSERVED- AND TRUE-SCORE EQUATING PROCEDURES

Martha L. Stocking
Daniel R. Eignor
Linda L. Cook

Educational Testing Service
Princeton, New Jersey
August 1988

Factors Affecting the Sample Invariant Properties of Linear and Curvilinear Observed- and True-Score Equating Procedures[1,2]

Martha L. Stocking[3]

Daniel R. Eignor

Linda L. Cook

Educational Testing Service

Princeton, New Jersey

---

[1]An earlier version of this paper was presented at the annual meeting of the American Educational Research Association, New Orleans, 1988.

[2]This study was supported by Educational Testing Service through Program Research Planning Council funding.

[3]The authors would like to recognize Maxine Kingston and Nancy Wright for programming and data preparation assistance and Charles Lewis, Robert Mislevy, Ledyard Tucker, Neil Dorans, Marilyn Wingersky, and Ida Lawrence for psychometric advice.

Abstract

A sequence of simulations was carried out to aid in the diagnosis and interpretation of equating differences found between random and matched (nonrandom) samples for four commonly used equating procedures: Tucker linear observed-score equating, Levine equally reliable linear observed-score equating, Equipercentile curvilinear observed-score equating, and IRT curvilinear true-score equating. The results support the prediction based on theoretical grounds that observed-score equating methods are more affected by sample variation than are true-score equating methods. These results further suggest that matching equating samples on the basis of fallible measures of ability may not be advisable for any conventional equating method except the Tucker method. In addition, the results support a particular hypothesis about IRT equating, suggesting that the use of matched samples cannot be recommended for this equating method either.

Factors Affecting the Sample Invariant Properties of Linear and Curvilinear Observed- and True-Score Equating Procedures

INTRODUCTION

For several decades, psychometricians have discussed and debated whether or not linear observed-score equating procedures such as the Tucker equating model (see Angoff, 1971) can provide invariant results when new and old form samples used in the equating differ in ability level. Levine (1955) developed a linear true-score equating model that was deemed to be more robust to differences in ability level of old and new form samples than the Tucker method. In the 1980's, IRT true-score equating (see Lord, 1980) won many advocates because of its claim to provide sample invariant equating results, provided the IRT model used fit the data and item parameters were adequately estimated. In the past few years, a number of studies have been performed to investigate the sample invariant properties of linear and IRT equating procedures (for example, Angoff & Cowell, 1986; Kingston, Leary, & Wightman, 1985; Cook, Eignor, & Taft, 1988); these studies have been reviewed and contrasted in a recent paper by Cook and Petersen (1987).

Lawrence and Dorans (1988) recently provided information addressing the sample invariant properties of Tucker and Levine linear equating and Equipercentile through an anchor test (Design V in Angoff, 1971) and three parameter logistic (3-PL) model IRT curvilinear equating in the context of equating the Scholastic Aptitude Test (SAT). Because the study to be described in this paper may be viewed in certain ways as an extension of the Lawrence and Dorans study, some of the details of the standard SAT data collection design and the matching process employed by Lawrence and Dorans in their study will be reviewed before results of the Lawrence and Dorans study will be discussed.

Figure 1 depicts the basic SAT equating data collection design, which essentially represents an equating design linking the new form, labelled NEW, to two old forms OLD1 and OLD2. The specific old forms to be used in the equating are established in the SAT braiding plan (Angoff, 1974); in general, the populations taking forms NEW and OLD1 will be populations of similar ability (data for form OLD1 will have been collected at the same administration during a previous year as form NEW), while the group of examinees taking form OLD2 will represent either a more or less able candidate population (data for form

OLD2 will have been collected at a different administration during a previous year than form NEW). Form NEW is linked to OLD1 via one anchor test (EQ1) and to OLD2 via another anchor test (EQ2). Typically, the average of the anchor equatings to the two old forms is taken as the operational conversion for the new form.

In the Lawrence and Dorans (1988) study, the authors focused on the equating of NEW to form OLD2, and in addition to performing the usual linear, Equipercentile through an anchor test, and 3-PL IRT equatings based on new and old form random samples that differ considerably in ability, matched sample equatings were alsc performed. In the matched sample equating of NEW to OLD2, the sample taking OLD2 (sample 4 in Figure 1) is chosen in a non-random fashion so that the old form distribution of scores on the anchor test (EQ2) matches the observed-score distribution of the new form equating sample (sample 2). Thus, while the observed-score distribution for the new form sample is the naturally occurring distribution, the observed-score distribution for the old form sample is altered under matched sample conditions to be similar to that of the new form sample. This matching procedure is seen as a means for controlling for the possible effects of ability level differences on equating

results. Lawrence and Dorans were then able to compare the random sample and matched sample equating results to determine which linear and curvilinear equating procedures provided the most and least invariant results.

Insert Figure 1 about here

Lawrence and Dorans (1988) studied and compared random and matched sample linear (Tucker and Levine), Equipercentile through an anchor test, and 3-PL IRT equatings (of NEW to OLD2) for nine forms of SAT-Mathematical and six forms of SAT-Verbal. The equating results, particularly scaled score means produced by the equating methods, revealed that the IRT true-score equating method was less robust to differences in group ability than expected, i.e., equating results for this method differed between the matched and unmatched (random) conditions. The Levine and Equipercentile through an anchor test equating results also differed considerably in certain equatings studied across the matched and random conditions. Interestingly, the Tucker observed-score equatings appeared more invariant across the matched and unmatched samples than any of the other methods. This was particularly true for the SAT-Mathematical equatings

studied, where there was little to no variation in scaled score means produced by the Tucker equating across the matched and random conditions. For SAT-Verbal, some variation in scaled score means resulting from the Tucker equatings was observed, but the sizes of the differences between the matched and random conditions was always less for Tucker than for other procedures. Further, while the four equating procedures frequently produced differing scaled score means under the random sample conditions, use of the anchor test as a direct selection variable for matching purposes produced a convergence of scaled score means across the four equating procedures. Lawrence and Dorans offered possible explanations for differences in equating results for all the procedures studied. Certain of these explanations, particularly the explanation for the IRT results, will be discussed later in this paper.

Consistency of equating results, and particularly scaled score means, across random and matched sample conditions was used as the criterion in the Lawrence and Dorans study. One potential problem with using consistency as the criterion is that consistent equating results may be disparate from the "true" equating results, were they known. In other words, the consistent Tucker equating results might have been more

disparate from the "true" equating results in the Lawrence and Dorans study than the inconsistent Levine or IRT equatings. Knowledge of "true" equating results suggests the need for a simulation study.

One recent simulation study supplied some useful results when considering the lack of invariance of the 3-PL IRT equatings. Stocking and Eignor (1986) showed that differences around one standard deviation between IRT equating sample mean abilities can have substantial effect (a five scaled score point difference) on the SAT mean scaled score when compared to results for samples not differing in mean ability and to "true" results. However, most of the random and matched sample equatings studied by Lawrence and Dorans (1988) showed as great or greater differences in score means as the Stocking and Eignor (1986) study although there were smaller differences in sample mean abilities. Hence the differences or lack of invariance of the 3-PL IRT equating results in the Lawrence and Dorans (1988) study suggests the design of a simulation study where more variables can be studied than simply ability level differences.

The goal of the present study was to develop a general simulation model and then perform a sequence of simulations and subsequent equatings based on the model that would address

specific issues in the application of both conventional and IRT-based equating methodologies, many of which were brought out in the Lawrence and Dorans (1988) study. More specifically, the purpose of the study was to investigate, using a sequence of simulations, the impact on four equating procedures of: 1) differences in abilities of samples used for equating, both when each examinee has complete data (an unrealistic setting) and also in the presence of missing data (a more realistic setting); 2) subsequent matching of samples on an infallible measure of ability (an unrealistic setting); and 3) subsequent matching of samples on a fallible measure of ability (a more realistic setting).

THE STUDY DESIGN

The Definition of True Item and Person Parameters

For the sequence of simulations described in this paper, true item and person parameters are required. They could, of course, be invented. It is more realistic, however, to use existing parameter estimates, but treat them as if they were true. It seems reasonable to assume that such a definition of truth captures at least some of the predominant features of actual data, such as the spread of abilities and item difficulties. For this purpose, the results of a LOGIST

calibration (Wingersky, Barton, & Lord, 1982) of a single 85-item SAT-Verbal test form (administered in two separately timed sections) plus a 45-item associated anchor test or equating section were used as the true item parameters. Descriptive statistics for these true item parameters are shown in Table 1.

Insert Table 1 about here

True person parameters were defined to be the ability estimates obtained when a sample of N = 3018 real examinees took the Verbal form and its associated equating section. Two population distributions of true ability were defined for this study. The first was defined to be exactly like the distribution of true person parameters, with mean true ability of -.02 and standard deviation of true ability equal to 1.07. A second population was defined to be less able, with mean true ability of -.35, and the same standard deviation.

For the purposes of this study, a total of six independent samples of size N = 3000 were drawn, as follows:

Sample	Drawn from Population	Sample Mean Ability	Sample Standard Deviation of Ability
1	1	-.01	1.06
2	1	-.03	1.08
3	1	-.02	1.06
4	1	.01	1.08
5	2	-.37	1.06
6[1]	2	-.06	1.08

The Generation of Complete Response Data

Two types of response data were generated for each simulated examinee (simulee). In this section, we discuss the generation of complete response strings; in a subsequent section, we describe the incorporation of missing data.

To generate responses to an item for a simulee, the simulee's true ability and the item's true 3-PL parameters are used to generate the model predicted probability of a correct response (see Lord, 1980). A random number is then selected from a uniform [0,1] distribution and compared to this model probability. If the random number is less than the modeled probability, the simulee is assigned a correct response to the item; if the random number is greater than the modeled probability, the simulee is assigned an incorrect response.

---

[1]Sample 6 was matched to Sample 2 using the observed formula-score distribution of Sample 2 on the anchor test.

This response string may be referred to as the true model-generated response string. It represents what the model says about examinee behavior for every item.

Models for Missing Response Data

Real examinees rarely have complete data. Data can be absent from a response string for at least two reasons. The examinee may not have had time to examine all test items, and therefore fails to respond to a block of items at the end of a test. This type of missing data is referred to as 'not-reached'. A second type of missing data occurs, particularly in formula scored tests, where an examinee may decide to omit an item because the examinee thinks that she/he can only respond at random. For whatever reason responses are missing, it seems most likely that the existence and patterns of missing data in response strings may be a function of the ability the test is designed to measure. This clearly violates the assumptions of the 3-PL model, and will almost certainly have some effect on calibration and equating results. It seems reasonable to attempt to incorporate this type of examinee behavior as one of the aspects to be studied in these simulations.

The mathematical modeling of missing responses is a complex and difficult process involving assumptions about the behavior

of examinees that may be difficult to test. This is clearly beyond the scope of the present paper. It is possible, however, to develop empirically-based models of missing data that make up for what they lack in generality by their close resemblance to real SAT data. It is important to note that, because the models proposed below are empirically based, they favor no particular treatment of missing data as incorporated into specific calibration procedures.

An Empirically-Based Model of Speededness

We wish to model speededness as a function of ability. To do this we need the actual item responses from each real examinee included in the calibration that produced our true item and person parameters. We can call these data the true response strings. We also need the true ability for each real examinee. Using the true response strings and true ability, we build a model of speededness only once, in advance of the simulations, for each separately timed test section. For each quintile of the distribution of true ability, we determine the cumulative distribution of the number of items reached for all examinees in the quintile. These conditional distributions will differ by ability level, and collectively they constitute our empirically-

based model. To incorporate this model in subsequent simulations, we proceed as follows:

1) Find the quintile into which a simulee's true ability falls.

2) Generate a random number between zero and one.

3) In the correct conditional distribution, find the cumulant that most closely matches the random number.

4) Find the corresponding number of items reached.

5) Assume that subsequent items are not reached for a simulee, and code 3's (the LOGIST code for not reached items) in the remainder of the model-generated response string for this simulee.

Figures 2, 3, and 4 show these empirically-based models separately for each separately timed section. In each of these figures, the frequency distribution of true abilities is plotted upside down; values of these proportional frequencies must be read from the right-hand vertical scale. This frequency distribution is divided into quintiles by the dotted vertical lines. In each figure, a solid vertical line is plotted at the midpoint of each quintile to serve as the x-axis for the cumulative conditional distributions, which are plotted sideways. The conditional distributions for each quintile are

the cumulative proportions of the individuals falling in that quintile who reached a specified proportion of the items in that section. Values for the specified proportion must be read from the left-hand vertical scale.

Insert Figures 2, 3, 4 about here

Although crude, these figures do demonstrate that this empirically-based model incorporates the number of items reached as a function of ability. For each separately timed section, there is a noticeable increase in the proportions of individuals completing more of the test as one looks across the quintiles from the lowest to the highest quintile.

An Empirically-Based Model of Omits

We assume the omitting behavior is a function of the ability to be measured by the test. We also assume an additional complexity -- that omitting also depends upon whether an examinee thinks she/he will get an item correct or incorrect. We need the same data as before, that is, the true response strings and true abilities for real examinees included in the calibration that produced our true item and person parameters. We also need additional data, that is, the true model-generated

response strings for the same examinees. This latter response string represents what the model predicts for each item for an examinee.

For each item, we construct two sub-models. The first is for those individuals whose model-predicted response was correct; we take this to indicate that the examinee thought she/he would get the item right. The second is for those individuals whose model-predicted response was incorrect; we take this to indicate that the examinee thought she/he would get the item wrong. For each sub-model, for each quintile of the distribution of true ability, we compute the proportions of examinees who omit the item in the true response strings. We construct these models for each item only once, using our true item and person parameters, true response strings, and true model-generated response strings. To incorporate these models in subsequent simulations, we proceed as follows:

1) For a true simulee ability, determine the model-generated response.

2) For the corresponding sub-model, find the corresponding quintile in the correct ability distribution.

3) Generate a random number between zero and one.

4) If the random number is less than the proportion of omits observed in the true response string, change the response in the simulee's model-generated response string to an omit. If the random number is greater than the proportion, do not change the response.

The empirically-based models of omitting behavior are shown for a few selected items in Figure 5. There are two plots for each item -- one for those examinees whose model generated responses indicated that they would respond incorrectly, and a second for those examinees whose model generated responses indicated that they would respond correctly. For each of these plots, the frequency distribution of true abilities for those examinees with the appropriate model-generated response is plotted upside down on the horizontal axis, with vertical bars marking off the quintiles. Actual values for this frequency distribution must be read from the bottom vertical axis. Above the horizontal axis in each figure, the proportion of individuals in a quintile whose true response strings indicated an omit are plotted with a cross at the midpoint of a quintile. These proportions are to be read from the top vertical axis.

Insert Figure 5 about here

There is some variation in omitting rates among the items displayed in Figure 5. Items 15, 55, 95 and 99 are hard items with more than 1000 omits (33%) in the full sample. Items 1 and 16 are easy items with fewer than 10 omits (.33%) in the full sample. Items 50 and 91 are items of middle difficulty; the rates of omitting in the true response strings are moderate. The following table gives the true parameters for these items.

Item Number	a	b	c	Number of Omits in full sample
15	.9	2.4	.18	>1000
55	.4	2.6	.13	>1000
95	1.0	1.4	.25	>1000
99	1.0	2.0	26	>1000
1	.3	-3.7	.12	<10
16	.6	-2.8	.12	<10
50	1.2	.0	.23	538
91	.8	.0	.10	270

Looking at these plots leads to a number of general conclusions. First, examinees who are modeled to get an item right tend to omit less frequently than those modeled to get an item wrong. This trend is most marked for those in the highest quintile of their respective ability distributions. Second, the rate of omitting is usually higher for lower ability, regardless

of the modeled response. This latter trend seems most consistent for those modeled to answer an item correctly. This model, then, reflects the two aspects we had hoped to incorporate, namely that omitting behavior is a function of ability and also a function of whether an examinee thinks that she/he will respond correctly.

The Design of the Calibrations and Equatings

The simulated responses from the six samples of simulees to the test form and equating section were combined into five concurrent LOGIST runs, each representing an experimental condition. The design of each LOGIST run was the same, and patterns in form the usual SAT data collection design presented in Figure 1.

Total Test or Anchor Test

	NEW	EQ1	EQ2	OLD1	OLD2
Sample 1	x	x			
Sample 2	x		x		
Sample 3		x		x	
Sample Y (Y=4,5, or 6)			x		x

Sample 1 was administered the new form and one anchor test (EQ1), Sample 2 was administered the new form and another anchor test (EQ2), Sample 3 was administered the first anchor test (EQ1) and an old form (OLD1), and a final sample (either Sample

4, 5 or 6) was administered the second anchor test (EQ2) and another old form (OLD2). All test forms (NEW, OLD1, OLD2) had identical true parameters, and all anchor tests (EQ1 and EQ2) had identical true item parameters.

From the item parameter estimates derived from each of the LOGIST runs or from the observed-score data for the samples used in the runs, the new form was equated to each old form using the Tucker, Levine, Equipercentile through an anchor test, and 3-PL IRT equating methods. The two equatings were also averaged to produce a final equating. All old forms were placed on the SAT 200 to 800 scaled score metric by the nonlinear equating originally derived for the SAT-Verbal form that serves as the source of the true item and person parameters. Projected scaled score means and standard deviations were computed for each single equating and each average using a sample of over 90,000 examinees who took that SAT-Verbal form at its initial equating administration.

The Scaling of Calibration Results

Many of the comparisons made in this study involve the estimated parameters obtained from separate LOGIST calibrations. However, each calibration will have results reported on a different metric, since LOGIST determines the reporting metric

by standardizing the ability estimates within a calibration. Therefore, the estimated parameters must all be placed on some common metric before such comparisons can be achieved.

The metric of the true item and person parameters was chosen as the common metric within which to compare parameter estimates. The parameter estimates from each LOGIST calibration were transformed to this common metric by the characteristic curve transformation method of Stocking and Lord (1983). The transformations were based on the parameter estimates from each calibration and the true parameters for the 130 items (85 test items plus 45 anchor test items) taken by Sample 1.

The Experimental Conditions

The series of simulations were designed to study five experimental conditions, shown in the following table, which contains a letter for each experimental condition:

	True Ability Distribution		
	Equivalent	Unequal	Equivalent by Matching
Complete data	A	B	-
Missing data	C	D	E

The data for all samples in a LOGIST run were either complete (conditions A and B) or contained missing data (conditions C, D, and E). The final samples taking EQ2 and OLD2, Samples 4-6,

were drawn in the following fashion. Sample 4 was drawn randomly from the same population as the other samples (conditions A and C); Sample 5 was drawn randomly from the lower ability population (conditions B and D); and Sample 6 was drawn from the lower ability population to match the distribution of observed formula scores obtained by Sample 2 on EQ2 (condition E).

Condition A, Complete Data and Equivalent Samples, is a benchmark condition in that, while unlikely to be realized in practice, it represents the best circumstances for any equating method. Condition B, Complete Data and Unequal Samples, provides for the exploration of the effects of different sample abilities while still maintaining the ideal situation of complete data for all simulees. This condition replicates the conditions of the Stocking and Eignor (1986) study, described in the introduction. Condition C, Missing Data and Equivalent Samples, is a more realistic condition in that samples now incorporate missing data. In this condition, however, samples have been chosen to be equivalent on the basis of an infallible criterion. Condition D, Missing Data and Unequal Samples, represents what is typically obtained in an SAT equating of NEW to OLD2 in the absence of any further data manipulation.

Condition E, Missing Data and Matched Samples, represents the matching procedure employed in the Lawrence and Dorans (1988) study; that is, matching samples on the basis of a fallible criterion in an attempt to achieve the ideal condition of equivalent samples.

RESULTS AND DISCUSSION

Calibration Results

Tables 2 through 6 contain descriptive statistics for the parameter estimates from each LOGIST calibration representing an experimental condition. In each table, the statistics for the item parameter estimates are given separately by test form or section. Statistics are also given for both the estimated and true abilities from each sample of simulees used in the calibration. These tables will be helpful in understanding some of the phenomena exhibited in the equating results.

Insert Tables 2, 3, 4, 5, and 6 about here

Equating Results

The focus of this study has been on the effect of the various experimental conditions on a number of different linear and curvilinear observed and true-score equating procedures.

For convenience, we divide the discussion of these equating results into two parts. In the first part, we examine the information available from this study relevant to a particular phenomenon observed by Lawrence and Dorans (1988) in their IRT equating results. This discussion is focused only on experimental condition D, Missing Data and Unequal Samples, and experimental condition E, Missing Data and Matched Samples. In the second part, we consider the results for all equating procedures across all experimental conditions.

An Exploration of the Lewis Hypothesis

Lawrence and Dorans (1988) observed that when the "matched" sample is more able than the "random" sample, i. e., Sample 6 is more able than Sample 5, the mean estimated item difficulty for OLD2 is higher when the estimates are obtained from Sample 6 than when obtained from Sample 5. When this is true, it automatically follows that the mean scaled score for NEW based on the matched sample calibration is lower than that based on the random-and-unequal sample calibration.

Charles Lewis (personal communication to Dorans, 1987) hypothesized the following circumstances to explain the difference in mean estimated item difficulties between the

random-and-unequal and matched conditions (experimental conditions D and E in the context of the current study):

1) Selecting Sample 6 from the same population as Sample 5 (a lower ability population) to match Sample 2 on the basis of observed scores on EQ2 will produce a sample of higher true ability than Sample 5, but not as high as the mean true ability for Sample 2. Given this level of true ability, the Sample 6 simulees will also have somewhat higher than expected observed scores on EQ2 (relative to Sample 5), corresponding to positive mean error scores in classical test theory.

2) The items in EQ2 will appear easier for Sample 6 than for Sample 2 because of the positive errors. LOGIST will try to reconcile these two sources of information about EQ2 items by estimating Sample 6 simulees to be more able than they actually are until the regressions of item score on estimated ability coincide for the two samples.

3) Items in OLD2 are also responded to by simulees in Sample 6, and by no other sample. If the simulees in Sample 6 are thought to be more able than they actually are, then their estimated abilities will be shifted to the right on the ability metric. The values of the estimated difficulties for items in OLD2 will be relative to the estimated abilities for Sample 6,

and since these abilities are shifted to the right, the estimated difficulties will be also, making these items appear harder than they actually are, relative to items in other forms.

4) In experimental condition D (Missing Data, Unequal Samples) of the current study, none of the distortions described above should occur. Thus the estimated difficulties for EQ2 for the two LOGIST calibrations should be approximately the same, while the estimated difficulties for the items in OLD2 arising from the matched sample condition (E) should be systematically greater than the corresponding difficulties for the random-and-unequal sample condition (D).

Lawrence and Dorans (1988) presented a table of average values for estimated item parameters for one of the SAT-Mathematical forms they studied under both experimental conditions. As in the case described above, the old form sample obtained by the matching process was more able than the randomly selected old form sample. The average difficulty for the old form affected by the change in sampling is about .08 higher under the matched sampling condition than under the random sampling condition, which supports the Lewis hypothesis.

Table 7 presents the same type of information as presented by Lawrence and Dorans, but for the current simulation.

However, in addition to the average values of item parameter estimates for the random-and-unequal case (D) and matched case (E), the same information is also presented for the Missing Data, Equivalent Samples condition (C), a condition that is equivalent to matching on an infallible criterion. As noted earlier, it is the results of this latter condition that the matching process is employed to achieve.

Insert Table 7 about here

Looking at the columns for item difficulty, we see that the average difficulty for the Missing Data, Matched Samples condition is .07 higher than the average difficulty for the Missing Data, Unequal Samples condition. In addition, there is little, if any, difference between the average difficulties for the other sections involved in the concurrent calibration. Differences between the averages of other item parameters are also small. These results replicate the Lawrence and Dorans (1988) results and support the Lewis hypothesis.

Perhaps even more notable, however, is the comparison of these two conditions with the "ideal" condition: Missing Data, Equivalent Samples. For the test forms and equating sections

not affected by the sample selection, there are few, if any, differences among the averages of the item parameter estimates across all three conditions. There is a change of only .01 in average item difficulty for OLD2, compared to the ideal condition, when there are true differences in ability. Matching samples on fallible criteria produces a much larger difference (.08) in average estimated difficulty. This suggests that such matching may introduce undesirable distortions in estimated item difficulties.

A more detailed comparison of results from the unequal samples and matched samples conditions is shown in Figure 6. Each page of this multipage figure shows a scatterplot (top) and residuals (bottom) for the item parameter estimates for a particular test section. In all scatterplots, the matched condition results are on the vertical axis and the random-and-unequal sample condition results are on the horizontal axis. All residual plots are formed by subtracting the unequal sample condition results from the matched sample condition results.

Insert Figure 6 about here

For the New form (NEW), EQ1, and OLD1, there are only a few items whose results do not lie exactly on the 45-degree line. These items are different because in one run their c's were fixed at COMC (see Wingersky, Barton, & Lord, 1982) and in the other run they were not. For EQ2, there is more scatter of the estimates around the 45-degree line, and the plot of item discriminations shows that the discriminations are slightly higher in the random condition, confirming the differences between the means in Table 1. For OLD2, there is even more scatter for all three item parameter estimates than seen for EQ2. The plot of the item difficulties shows that the estimates under the matched condition are generally slightly, but systematically, higher for almost all item difficulties.

To examine the same type of information for real data, as opposed to the simulated data developed for this study, a particular SAT-Verbal form studied by Lawrence and Dorans (1988) was selected. The form was chosen because the reported differences showed that the average ability for the lower ability sample taking one old form was about 1/3 of a standard deviation below that for the new form. This resembles the simulated conditions of the current study.

Figure 7 shows the information for this form analogous to that shown for the simulated data in Figure 6. The calibration design for the chosen form was exactly the same as in the simulation, but in contrast to the simulation, each test form and equating section for the real-data calibration differed from each other. As in the simulated results, item parameter estimates for NEW, EQ1 and OLD1 were not affected by the sample selection. Estimates for EQ2 and OLD2 were affected, and in much the same way as the simulated results. The item difficulty estimates for OLD2 are slightly, but systematically, higher in the matched condition. These results, as do the simulation results, provide further evidence in support of the Lewis hypothesis.

Insert Figure 7 about here

Table 7 and Figures 6 and 7 suggest that if IRT equating is to be used, then the matching of samples based on a fallible criterion is not recommended. This selection produces results that differ more from the ideal condition of selection on an infallible criterion than do the results based on the use of samples that are unequal in true ability. At the same time,

this selection introduces an undesirable bias in the estimates of item difficulty for the old form.

Equating Results for All Methods and All Conditions

Table 8 shows the projected scaled score means and standard deviations for all individual equatings performed and for the averages. Figure 8 plots the projected scaled score means for the individual equatings (not the averages). The left side of this figure gives the results of the equatings of the New Form to Old Form 1, and the right side gives the results for the equatings of the New Form to Old Form 2. The experimental conditions are positioned along the horizontal axis. The projected scaled score means are read from the vertical axis. For each experimental condition, the projected scaled score means are labeled with a T for Tucker, L for Levine, E for Equipercentile, and I for IRT equating. The points for a particular equating method are connected with straight lines to make the plots easier to read.

Insert Table 8 and Figure 8 about here

Both Table 8 and Figure 8 show that the differences among projected mean scaled scores are generally small, but with a few exceptions to be discussed later. This is not surprising since all test forms have the same true item parameters in this simulation; only samples have been changed. Thus equating the NEW to OLD1 or to OLD2 is equivalent to equating a test to itself, using identical anchor test sections. The importance of these small differences is not possible to judge since approximate standard errors have not been developed for all methods (i.e., the IRT standard errors have not been developed to date).

To evaluate these results, it seems useful to compare the results of each equating method across experimental conditions to its own value in the "benchmark" condition. This condition, shown to the far left of each subplot, is the one in which data are complete for each simulee and all samples of simulees are drawn from the same ability distribution.

New Form Equated to Old Form 1

Conventional equating methods for equating NEW to OLD1 are not affected by different samples taking OLD2 since these samples do not enter into the equating. Thus, the equated means for the conventional methods are identical for conditions

involving complete data (A and B), and also identical, but different, for conditions involving missing data (C, D, and E). In contrast, since all test forms are calibrated concurrently, IRT equating results vary slightly across conditions in which the samples taking the other old form vary.

All equating methods are affected by the presence of missing data in both the NEW and OLD1 samples (conditions C vs. A and conditions D vs. B), although IRT equating is less affected than conventional methods. The kind of missing data modeled here, in which both the number of items reached and omitted are functions of ability, tends to make all simulees appear slightly less able and the tests to appear slightly harder. In the IRT case, the new form is harder than the old form when there is complete data (see Table 2 or Table 3). When missing data is introduced, both test forms are harder, but differentially so, and the old form becomes even easier than the new form (see Table 4 or Table 5). Thus the new form scaled score mean is raised by introducing missing data.

For the IRT equatings, all other effects are not explainable on the basis of means of estimated item parameters, but may be explainable by slight changes in the distributions of

item parameter estimates due to what is most likely sampling variability.

If comparison with respective benchmark conditions is a reasonable criteria, then IRT shows the least variation across conditions studied.

New Form Equated to Old Form 2

These equatings, shown in the right-hand subplot of Figure 8, are the interesting ones -- by design they are most affected by the experimental conditions. As seen in Figure 8 and also in Table 8, the benchmark conditions for all equating methods are different from the benchmark conditions for the equating of NEW to OLD1. The IRT benchmark conditions are most different -- over two scaled score points; the Equipercentile benchmark conditions are least different -- less than a tenth of a scaled score point.

Perhaps the most striking aspect of these equatings is the sensitivity of observed-score equating methods to differences in true sample ability. The introduction of unequal samples, whether in the complete data situation (conditions B and A) or in the missing data conditions (conditions D and C) has the largest impact on Tucker equating, and less but substantial

impact on Equipercentile equating. Of the remaining two methods, Levine equating is more affected than the IRT equating.

As in the OLD1 equatings, the introduction of missing data (conditions C vs. A and conditions D vs. B) also impacts the projected means, making them slightly higher for all equating methods. The explanation for the IRT results offered previously for the equating of NEW to OLD1 seems to hold here also.

The Lewis hypothesis is again demonstrated by the slight decrease in the projected mean for IRT equating from the random-and-unequal sample condition (D) to the matched sample condition (E). Tucker and Levine equatings are identical, as they must be, under matched sampling conditions, and the Equipercentile equating is close to them.

If the benchmark condition is used as a criterion, it seems clear that IRT equating varies least across all experimental conditions. If the Missing Data, Equivalent Samples condition (C) is a more practical criterion, in other missing data conditions (D and E), all equating methods except Tucker come closer to this criterion when random-and-unequal samples are used than when matched samples are used. The matching process appears to improve the Tucker method, while making the other methods worse.

These results suggest that if Levine, Equipercentile, or IRT equatings are to be used, more reasonable results are obtained using random-and-unequal samples. If Tucker equating is to be used, better results are obtained with matched samples than with random-and-unequal samples. However, if the decision concerning the choice of equating procedure is to be made after the sampling decision, then these results suggest that it is better to use the random-and-unequal sampling that typically occurs in SAT equating situations, and never select the Tucker method.

CONCLUSIONS

The conclusions reported in this study must be considered tentative since they are based on a single sequence of simulations, and will remain tentative until they are replicated by other studies. Further, the results should be examined from the viewpoint that response data were generated according to the 3-PL model, with some specific model violations introduced to incorporate missing data. These circumstances may favor the 3-PL IRT equating results. In addition, it is not possible to draw definitive conclusions about the importance of the equating differences until some other study produces estimates of standard errors for all equating methods studied.

With the above in mind, the following tentative conclusions may be offered based on the results of this study:

1. If IRT true-score equating procedures are to be employed, matching of samples based on a fallible criterion, such as an anchor test observed-score distribution, is not recommended. This selection produces results that differ more from the ideal condition of selection on an infallible criterion than do the results based on the use of samples of unequal ability. Such selection also introduces an undesirable bias in the estimates of item difficulty for the old form.
2. If Levine equally reliable or Equipercentile through an anchor test observed-score equating procedures are to be employed, more reasonable results are also obtained from use of samples of unequal ability and matching is not recommended. Only for Tucker equating are better results obtained when samples are matched on a fallible criterion.

Finally, it is reasonable to ask how the results of this study compare to the real data results observed in the Lawrence and Dorans (1988) study. Their study involved looking at only

conditions D and E of the equating of NEW to OLD2 in Figure 8; they were, however, able to observe the results for a number of different forms. The results of this study for conditions D and E of the equating of NEW to OLD2 are not totally inconsistent with the Lawrence and Dorans findings for SAT-Verbal, and, in fact, the results reported in this study closely correspond to the results for one of the forms studied by Lawrence and Dorans. The results from this study are somewhat inconsistent with the Lawrence and Dorans findings for SAT-Mathematical, where little variation was found across the Tucker results for conditions D and E of the equating of NEW to OLD2. Further investigations are presently being planned to attempt to reconcile the inconsistency of equating results that appear to exist for the Tucker method for SAT-Verbal and SAT-Mathematical.

References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational measurement (pp. 508-600). Washington, DC: American Council on Education.

Angoff, W. H. (1974). The College Board Admissions Testing Program: A technical report on research and development activities relating to the Scholastic Aptitude Test and Achievement Tests. New York: College Entrance Examination Board.

Angoff, W. H., & Cowell, W. R. (1986). An examination of the assumption that the equating of parallel forms is population-independent. Journal of Educational Measurement, 23, 327-345.

Cook, L. L., Eignor, D. R., & Taft, H. L. (1988). A comparative study of the effects of recency of instruction on the stability of IRT and conventional item parameter estimates. Journal of Educational Measurement, 25, 31-45.

Cook, L. L., & Petersen, N. S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. Applied Psychological Measurement, 11, 225-244.

Kingston, N., Leary, L., & Wightman, L. (1985). An exploratory study of the applicability of item response theory methods to the Graduate Management Admissions Test (RR-85-34). Princeton, NJ: Educational Testing Service.

Lawrence I. L., & Dorans, N. J. (1988). A comparison of observed score and true score equating methods for representative samples and samples matched on an anchor test (RR-88-23). Princeton, NJ: Educational Testing Service.

Levine, R. S. (1955). Equating the score scales of alternate forms administered to samples of different ability (RB-55-23). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Stocking, M. L., & Eignor, D. R. (1986). The impact of different ability distributions on IRT pre-equating (RR-86-49). Princeton, NJ: Educational Testing Service.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. Applied Psychological Measurement, 7, 201-210.

Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). LOGIST V user's guide. Princeton, NJ: Educational Testing Service.

Table 1

Summary Statistics for the True SAT-Verbal Item and Person Parameters

True Parameters	N	Mean	S.D.	Min	Max	Percentiles 10	25	50	75	90
True a, Total Test	85	0.83	0.30	0.22	1.67	0.43	0.63	0.84	1.02	1.15
True b, Total Test	85	0.15	1.48	-3.66	2.59	-1.86	-1.00	0.62	1.38	1.83
True c, Total Test	85	0.17	0.08	0.00	0.50	0.10	0.12	0.16	0.21	0.27
True a, Anchor Test	45	0.88	0.29	0.34	1.59	0.43	0.66	0.88	1.06	1.18
True b, Anchor Test	45	0.22	1.34	-2.68	2.13	-1.63	-0.76	0.49	1.37	1.76
True c, Anchor Test	45	0.17	0.08	0.00	0.41	0.08	0.11	0.15	0.23	0.26
True abilities	3018	-0.02	1.07	-6.93	3.74	-1.28	-0.67	-0.01	0.65	1.27

Table 2

Summary Statistics for the Estimated Item and Person Parameters Estimated for Condition A: Complete Data and Equivalent Samples. Also Summary Statistics for the True Abilities of Samples 1 through 4. This Condition is the Benchmark Condition.

					Percentiles				
Parameters	N	Mean	S.D.	Min	Max	10	25	50	75	90
Estimated a, NEW	85	0.83	0.31	0.20	1.64	0.40	0.62	0.84	1.03	1.22
Estimated b, NEW	85	0.14	1.43	-3.44	2.60	-1.88	-0.93	0.54	1.35	1.82
Estimated c, NEW	85	0.16	0.09	0.00	0.50	0.06	0.10	0.15	0.20	0.26
Estimated a, OLD1	85	0.82	0.30	0.19	1.64	0.41	0.61	0.81	1.02	1.17
Estimated b, OLD1	85	0.13	1.47	-3.81	2.35	-1.91	-1.05	0.55	1.32	1.85
Estimated c, OLD1	85	0.15	0.09	0.00	0.48	0.03	0.10	0.14	0.20	0.25
Estimated a, OLD2	85	0.81	0.31	0.24	1.73	0.45	0.60	0.77	0.98	1.18
Estimated b, OLD2	85	0.14	1.54	-4.29	2.64	-1.83	-0.96	0.60	1.39	1.93
Estimated c, OLD2	85	0.15	0.08	0.00	0.45	0.05	0.10	0.14	0.19	0.26
Estimated a, EQ1	45	0.87	0.29	0.30	1.73	0.45	0.69	0.88	1.02	1.19
Estimated b, EQ1	45	0.21	1.34	-2.30	2.18	-1.63	-0.83	0.43	1.38	1.88
Estimated c, EQ1	45	0.16	0.09	0.00	0.40	0.05	0.10	0.13	0.22	0.27
Estimated a, EQ2	45	0.87	0.30	0.32	1.73	0.42	0.73	0.86	1.03	1.27
Estimated b, EQ2	45	0.21	1.34	-2.56	2.21	-1.67	-0.84	0.43	1.41	1.75
Estimated c, EQ2	45	0.16	0.08	0.00	0.38	0.06	0.10	0.13	0.21	0.26
Est. abilities, S1	3000	0.01	1.08	-7.35	4.59	-1.26	-0.64	0.00	0.71	1.32
Est. abilities, S2	3000	-0.01	1.07	-6.31	3.78	-1.29	-0.63	-0.01	0.66	1.31
Est. abilities, S3	3000	-0.00	1.09	-7.35	3.77	-1.30	-0.66	-0.02	0.64	1.37
Est. abilities, S4	3000	0.03	1.12	-7.35	5.00	-1.27	-0.65	0.01	0.70	1.37
True abilities, S1	3000	-0.01	1.06	-6.93	3.74	-1.25	-0.64	-0.00	0.69	1.26
True abilities, S2	3000	-0.03	1.08	-6.93	3.56	-1.26	-0.63	-0.01	0.64	1.22
True abilities, S3	3000	-0.02	1.06	-6.45	3.51	-1.27	-0.66	-0.02	0.62	1.34
True abilities, S4	3000	0.01	1.08	-6.93	3.56	-1.26	-0.65	0.01	0.70	1.29

Table 3

Summary Statistics for the Estimated Item and Person Parameters Estimated for Condition B: Complete Data and Unequal Samples. Also Summary Statistics for the True Abilities of Samples 1, 2, 3, and 5.

						Percentiles				
Parameters	N	Mean	S.D.	Min	Max	10	25	50	75	90
Estimated a, NEW	85	0.83	0.31	0.20	1.64	0.40	0.62	0.84	1.03	1.23
Estimated b, NEW	85	0.14	1.48	-3.43	2.60	-1.82	-0.93	0.54	1.35	1.82
Estimated c, NEW	85	0.16	0.09	0.00	0.50	0.06	0.10	0.15	0.20	0.26
Estimated a, OLD1	85	0.82	0.30	0.19	1.64	0.41	0.61	0.81	1.01	1.17
Estimated b, OLD1	85	0.13	1.47	-3.81	2.35	-1.91	-1.06	0.55	1.32	1.85
Estimated c, OLD1	85	0.15	0.09	0.00	0.48	0.03	0.10	0.14	0.20	0.25
Estimated a, OLD2	85	0.84	0.32	0.23	1.72	0.42	0.63	0.81	1.05	1.21
Estimated b, OLD2	85	0.16	1.52	-3.66	2.71	-1.95	-0.97	0.57	1.42	1.87
Estimated c, OLD2	85	0.15	0.09	0.00	0.49	0.03	0.10	0.15	0.20	0.25
Estimated a, EQ1	45	0.86	0.29	0.30	1.72	0.45	0.69	0.88	1.02	1.19
Estimated b, EQ1	45	0.21	1.34	-2.80	2.18	-1.64	-0.83	0.42	1.38	1.88
Estimated c, EQ1	45	0.16	0.09	0.00	0.40	0.05	0.10	0.13	0.22	0.27
Estimated a, EQ2	45	0.86	0.30	0.31	1.72	0.43	0.69	0.86	1.02	1.26
Estimated b, EQ2	45	0.21	1.34	-2.61	2.38	-1.61	-0.77	0.40	1.35	1.89
Estimated c, EQ2	45	0.16	0.08	0.00	0.38	0.06	0.10	0.13	0.21	0.25
Est. abilities, S1	3000	0.01	1.08	-7.46	4.60	-1.26	-0.64	0.00	0.71	1.32
Est. abilities, S2	3000	-0.01	1.07	-6.45	3.81	-1.29	-0.63	-0.00	0.66	1.31
Est. abilities, S3	3000	-0.01	1.09	-7.46	3.77	-1.30	-0.67	-0.03	0.64	1.37
Est. abilities, S5	3000	-0.34	1.06	-7.46	3.74	-1.58	-0.96	-0.36	0.33	0.93
True abilities, S1	3000	-0.01	1.06	-6.93	3.74	-1.25	-0.64	-0.00	0.69	1.26
True abilities, S2	3000	-0.03	1.08	-6.93	3.56	-1.24	-0.63	-0.00	0.64	1.22
True abilities, S3	3000	-0.02	1.06	-6.45	3.51	-1.27	-0.66	-0.02	0.62	1.34
True abilities, S5	3000	-0.37	1.06	-7.26	3.40	-1.65	-1.01	-0.36	0.30	0.92

Table 4

Summary Statistics for the Estimated Item and Person Parameters Estimated for Condition C: Missing Data and Equivalent Samples. Also Summary Statistics for the True Abilities of Samples 1, 2, 3, and 4.

					Percentiles				
Parameters	N	Mean	S.D.	Min	Max	10	25	50	75	90
Estimated a, NEW	85	0.85	0.31	0.20	1.69	0.43	0.64	0.88	1.04	1.20
Estimated b, NEW	85	0.17	1.48	-3.38	3.59	-1.75	-1.03	0.48	1.38	1.89
Estimated c, NEW	85	0.16	0.08	0.00	0.49	0.08	0.12	0.16	0.20	0.24
Estimated a, OLD1	85	0.83	0.30	0.20	1.68	0.46	0.65	0.81	1.01	1.18
Estimated b, OLD1	85	0.15	1.47	-3.73	3.11	-1.76	-1.18	0.52	1.35	1.83
Estimated c, OLD1	85	0.16	0.08	0.00	0.47	0.04	0.12	0.14	0.20	0.25
Estimated a, OLD2	85	0.84	0.31	0.25	1.74	0.41	0.62	0.82	1.00	1.22
Estimated b, OLD2	85	0.16	1.51	-4.05	3.22	-1.79	-1.03	0.50	1.38	1.86
Estimated c, OLD2	85	0.16	0.08	0.00	0.44	0.06	0.12	0.16	0.20	0.26
Estimated a, EQ1	45	0.91	0.31	0.38	1.74	0.46	0.71	0.91	1.07	1.26
Estimated b, EQ1	45	0.24	1.36	-2.63	2.38	-1.61	-0.88	0.36	1.31	1.95
Estimated c, EQ1	45	0.16	0.07	0.00	0.35	0.09	0.12	0.14	0.21	0.26
Estimated a, EQ2	45	0.90	0.32	0.33	1.74	0.44	0.73	0.86	1.09	1.35
Estimated b, EQ2	45	0.25	1.37	-2.54	2.53	-1.66	-0.94	0.43	1.42	1.91
Estimated c, EQ2	45	0.16	0.07	0.00	0.33	0.08	0.12	0.15	0.21	0.25
Est. abilities, S1	3000	-0.12	1.06	-7.40	3.34	-1.37	-0.75	-0.14	0.55	1.19
Est. abilities, S2	3000	-0.14	1.05	-6.40	3.10	-1.38	-0.76	-0.13	0.52	1.15
Est. abilities, S3	3000	-0.13	1.06	-7.40	3.41	-1.39	-0.78	-0.15	0.52	1.26
Est. abilities, S4	3000	-0.10	1.09	-7.40	3.62	-1.35	-0.77	-0.12	0.57	1.24
True abilities, S1	3000	-0.01	1.06	-6.93	3.74	-1.25	-0.64	-0.00	0.69	1.26
True abilities, S2	3000	-0.03	1.08	-6.93	3.56	-1.24	-0.63	-0.00	0.64	1.22
True abilities, S3	3000	-0.02	1.06	-6.45	3.51	-1.27	-0.66	-0.02	0.62	1.34
True abilities, S4	3000	0.01	1.08	-6.93	3.56	-1.26	-0.65	0.01	0.70	1.29

Table 5

Summary Statistics for the Estimated Item and Person Parameters Estimated for Condition D: Missing Data and Unequal Samples. Also Summary Statistics for the True Abilities of Samples 1, 2, 3, and 5.

						Percentiles				
Parameters	N	Mean	S.D.	Min	Max	10	25	50	75	90
Estimated a, NEW	85	0.85	0.31	0.20	1.69	0.43	0.64	0.88	1.04	1.20
Estimated b, NEW	85	0.17	1.48	-3.38	3.59	-1.75	-1.03	0.48	1.38	1.89
Estimated c, NEW	85	0.16	0.08	0.00	0.49	0.08	0.12	0.16	0.20	0.24
Estimated a, OLD1	85	0.83	0.30	0.20	1.68	0.46	0.65	0.81	1.01	1.18
Estimated b, OLD1	85	0.15	1.48	-3.74	3.11	-1.76	-1.18	0.52	1.35	1.84
Estimated c, OLD1	85	0.16	0.08	0.00	0.47	0.04	0.12	0.14	0.20	0.25
Estimated a, OLD2	85	0.85	0.33	0.23	1.74	0.39	0.64	0.85	1.07	1.24
Estimated b, OLD2	85	0.17	1.50	-3.46	3.64	-1.78	-1.10	0.47	1.34	1.80
Estimated c, OLD2	85	0.16	0.08	0.00	0.46	0.04	0.12	0.15	0.21	0.24
Estimated a, EQ1	45	0.91	0.30	0.38	1.74	0.51	0.71	0.91	1.07	1.26
Estimated b, EQ1	45	0.25	1.36	-2.64	2.38	-1.62	-0.88	0.37	1.31	1.96
Estimated c, EQ1	45	0.16	0.07	0.00	0.35	0.09	0.12	0.15	0.22	0.26
Estimated a, EQ2	45	0.90	0.31	0.30	1.74	0.45	0.75	0.86	1.06	1.27
Estimated b, EQ2	45	0.24	1.37	-2.57	2.24	-1.66	-0.86	0.37	1.44	1.98
Estimated c, EQ2	45	0.16	0.08	0.00	0.35	0.06	0 12	0.15	0.21	0.25
Est. abilities, S1	3000	-0.12	1.06	-7.51	3.34	-1.37	-0.75	-0.14	0.55	1.20
Est. abilities, S2	3000	-0.14	1.05	-6.41	3.09	-1.38	-0.76	-0.13	0.52	1.16
Est. abilities, S3	3000	-0.13	1.06	-7.51	3.42	-1.39	-0.78	-0.15	0.52	1.26
Est. abilities, S5	3000	-0.46	1.04	-7.51	3.24	-1.68	-1.08	-0.49	0.19	0.81
True abilities, S1	3000	-0.01	1.06	-6.93	3.74	-1.25	-0.64	-0.00	0.69	1.26
True abilities, S2	3000	-0.03	1.08	-6.93	3.56	-1.24	-0.63	-0.00	0.64	1.22
True abilities, S3	3000	-0.02	1.06	-6.45	3.51	-1.27	-0.66	-0.02	0.62	1.34
True abilities, S5	3000	-0.37	1.06	-7.26	3.40	-1.65	-1.01	-0.36	0.30	0.92

Table 6

Summary Statistics for the Estimated Item and Person Parameters Estimated for Condition E: Missing Data and Matched Samples. Also Summary Statistics for the True Abilities of Samples 1, 2, 3, and 6.

						Percentiles				
Parameters	N	Mean	S.D.	Min	Max	10	25	50	75	90
Estimated a, NEW	85	0.85	0.31	0.20	1.69	0.44	0.64	0.88	1.03	1.20
Estimated b, NEW	85	0.17	1.48	-3.37	3.59	-1.74	-1.03	0.48	1.38	1.89
Estimated c, NEW	85	0.16	0.08	0.00	0.49	0.08	0.12	0.16	0.20	0.24
Estimated a, OLD1	85	0.83	0.30	0.20	1.68	0.46	0.65	0.81	1.01	1.18
Estimated b, OLD1	85	0.15	1.47	-3.72	3.10	-1.76	-1.18	0.52	1.35	1.84
Estimated c, OLD1	85	0.16	0.08	0.00	0.47	0.04	0.12	0.14	0.20	0.25
Estimated a, OLD2	85	0.87	0.32	0.24	1.74	0.45	0.65	0.86	1.08	1.26
Estimated b, OLD2	85	0.24	1.50	-3.09	4.40	-1.70	-0.97	0.46	1.42	1.97
Estimated c, OLD2	85	0.17	0.08	0.00	0.47	0.08	0.12	0.16	0.21	0.27
Estimated a, EQ1	45	0.91	0.31	0.38	1.74	0.46	0.71	0.91	1.07	1.26
Estimated b, EQ1	45	0.24	1.36	-2.62	2.38	-1.61	-0.88	0.37	1.31	1.95
Estimated c, EQ1	45	0.16	0.07	0.00	0.35	0.09	0.12	0.14	0.21	0.26
Estimated a, EQ2	45	0.88	0.31	0.27	1.74	0.45	0.72	0.85	1.08	1.31
Estimated b, EQ2	45	0.24	1.37	-2.42	2.39	-1.67	-0.82	0.51	1.45	2.02
Estimated c, EQ2	45	0.16	0.07	0.00	0.36	0.08	0.12	0.14	0.22	0.24
Est. abilities, S1	3000	-0.12	1.06	-7.38	3.34	-1.37	-0.75	-0.14	0.55	1.19
Est. abilities, S2	3000	-0.14	1.05	-6.53	3.12	-1.38	-0.76	-0.13	0.52	1.16
Est. abilities, S3	3000	-0.13	1.06	-7.38	3.42	-1.39	-0.78	-0.15	0.52	1.26
Est. abilities, S6	3000	-0.15	1.05	-7.38	3.05	-1.40	-0.79	-0.18	0.52	1.14
True abilities, S1	3000	-0.01	1.06	-6.93	3.74	-1.25	-0.64	-0.00	0.69	1.26
True abilities, S2	3000	-0.03	1.08	-6.93	3.56	-1.24	-0.63	-0.00	0.64	1.22
True abilities, S3	3000	-0.02	1.06	-6.45	3.51	-1.27	-0.66	-0.02	0.62	1.34
True abilities, S6	3000	-0.06	1.08	-7.26	3.23	-1.33	-0.69	-0.06	0.65	1.22

Table 7

Means and Standard Deviations of Item Parameter Estimates for the Three Missing-Data Conditions -- E = Equivalent Samples (Condition C), R = Random-and-Unequal Samples (Condition D), M = Matched Samples (Condition E)

		a				b				c			
Form		E	R	M	R-M	E	R	M	R-M	E	R	M	R-M
OLD2	$\bar{X}$	.84	.85	.87	-.02	.16	.17	.24	-.07	.16	.16	.17	-.01
	SD	.31	.33	.32		1.51	1.50	1.50		.08	.08	.08	
EQ2	$\bar{X}$	.90	.90	.88	.02	.25	.24	.24	.00	.16	.16	.16	.00
	SD	.32	.31	.31		1.37	1.37	1.37		.07	.08	.07	
NEW	$\bar{X}$	.85	.85	.85	.00	.17	.17	.17	.00	.16	.16	.16	.00
	SD	.31	.31	.31		1.48	1.48	1.48		.08	.08	.08	
OLD1	$\bar{X}$	.83	.83	.83	.00	.15	.15	.15	.00	.16	.16	.16	.00
	SD	.30	.30	.30		1.47	1.48	1.47		.08	.08	.08	
EQ1	$\bar{X}$	.91	.91	.91	.00	.24	.25	.24	.01	.16	.16	.16	.00
	SD	.31	.30	.31		1.36	1.36	1.36		.07	.07	.07	

Table 8: Projected Scaled Score Means and Standard Deviations for All Equating Methods and All Experimental Conditions

		Tucker					
	Condition	NEW to OLD1		NEW to OLD2		Average	
		Mean	S.D.	Mean	S.D.	Mean	S.D.
Benchmark	A	420.72	112.39	421.22	108.52	420.96	110.44
Compdata,uneq	B	420.72	112.39	414.90	106.31	417.80	109.34
Missdata,equal	C	422.10	111.14	421.71	109.14	421.89	110.13
Missdata,uneq	D	422.10	111.14	415.35	107.92	418.71	109.07
Missdata,matched	E	422.10	111.14	417.95	108.92	420.02	110.02

		Levine					
	Condition	NEW to OLD1		NEW to OLD2		Average	
		Mean	S.D.	Mean	S.D.	Mean	S.D.
Benchmark	A	420.89	112.30	420.79	107.55	420.83	109.91
Compdata,uneq	B	420.89	112.30	420.06	106.97	420.47	109.62
Missdata,equal	C	422.31	110.87	421.15	108.42	421.73	109.63
Missdata,uneq	D	422.31	110.87	420.42	108.01	421.36	109.43
Missdata,matched	E	422.31	110.87	417.95	108.92	420.13	109.88

		Equipercentile					
	Condition	NEW to OLD1		NEW to OLD2		Average	
		Mean	S.D.	Mean	S.D.	Mean	S.D.
Benchmark	A	420.74	112.77	420.82	107.85	420.81	110.24
Compdata,uneq	B	420.74	112.77	418.76	107.39	419.78	110.00
Missdata,equal	C	422.00	110.67	421.05	108.24	421.52	109.38
Missdata,uneq	D	422.00	110.67	419.04	108.02	420.52	109.28
Missdata,matched	E	422.00	110.67	417.82	108.93	419.90	109.72

		IRT					
	Condition	NEW to OLD1		NEW to OLD2		Average	
		Mean	S.D.	Mean	S.D.	Mean	S.D.
Benchmark	A	422.12	111.10	419.79	109.13	420.95	110.12
Compdata,uneq	B	422.35	110.99	419.70	109.56	420.76	110.27
Missdata,equal	C	422.52	110.37	420.46	108.94	421.49	109.65
Missdata,uneq	D	422.77	110.17	420.12	109.90	421.45	110.04
Missdata,matched	E	422.50	110.33	419.07	108.68	420.79	109.50

Figure 1. Data collection design for equating the SAT

	Total Test or Anchor Test				
	NEW	EQ1	EQ2	OLD1	OLD2
Sample 1	X	X			
Sample 2	X		X		
Sample 3		X		X	
Sample 4			X		X

Notes: An X denotes the specific total test and anchor test taken by a specific sample.

Samples 1 and 2 are random samples from the same total group.

Samples 1 and 3 are samples from different total groups that are similar in ability.

Samples 2 and 4 are samples from different total groups that are dissimilar in ability.

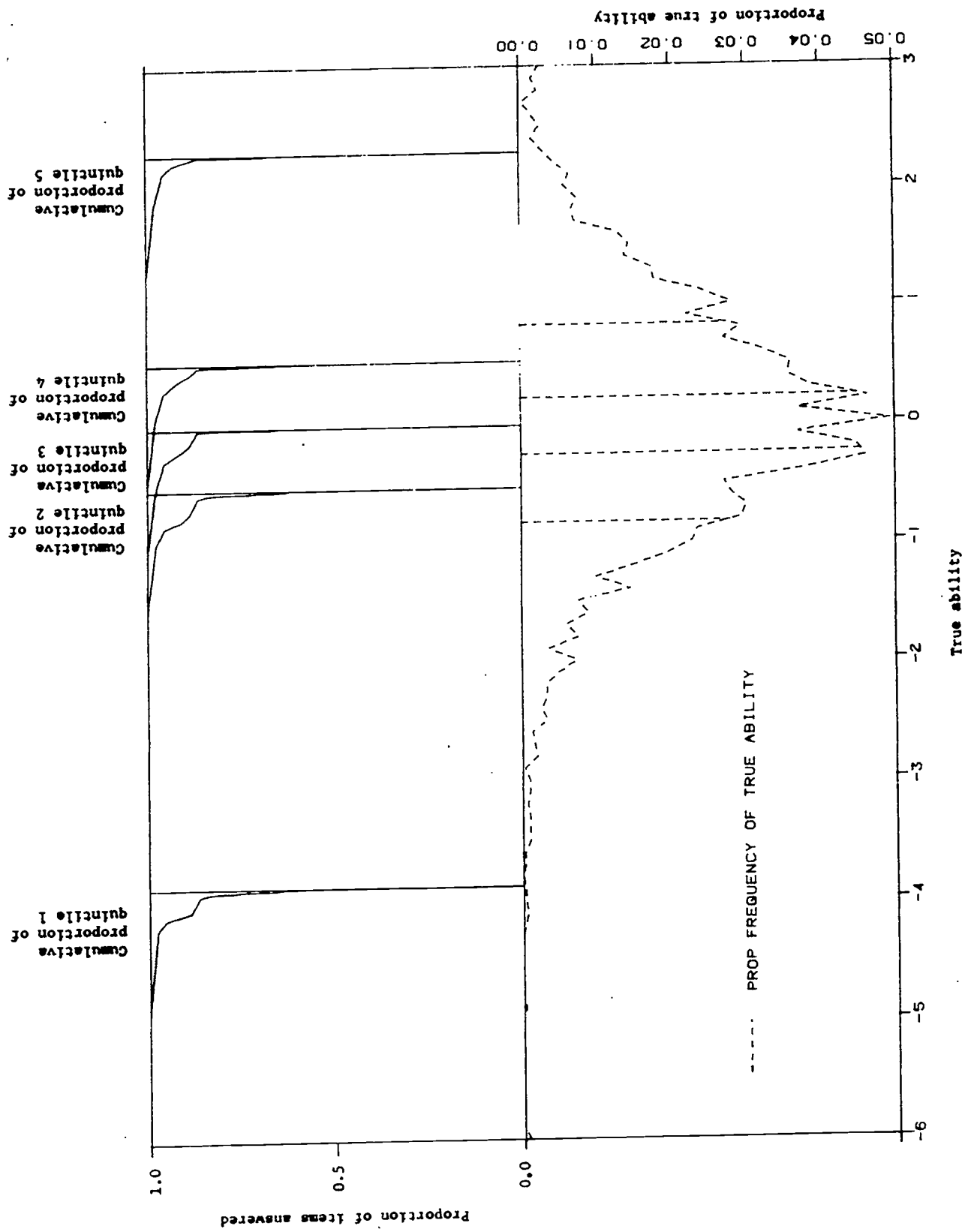
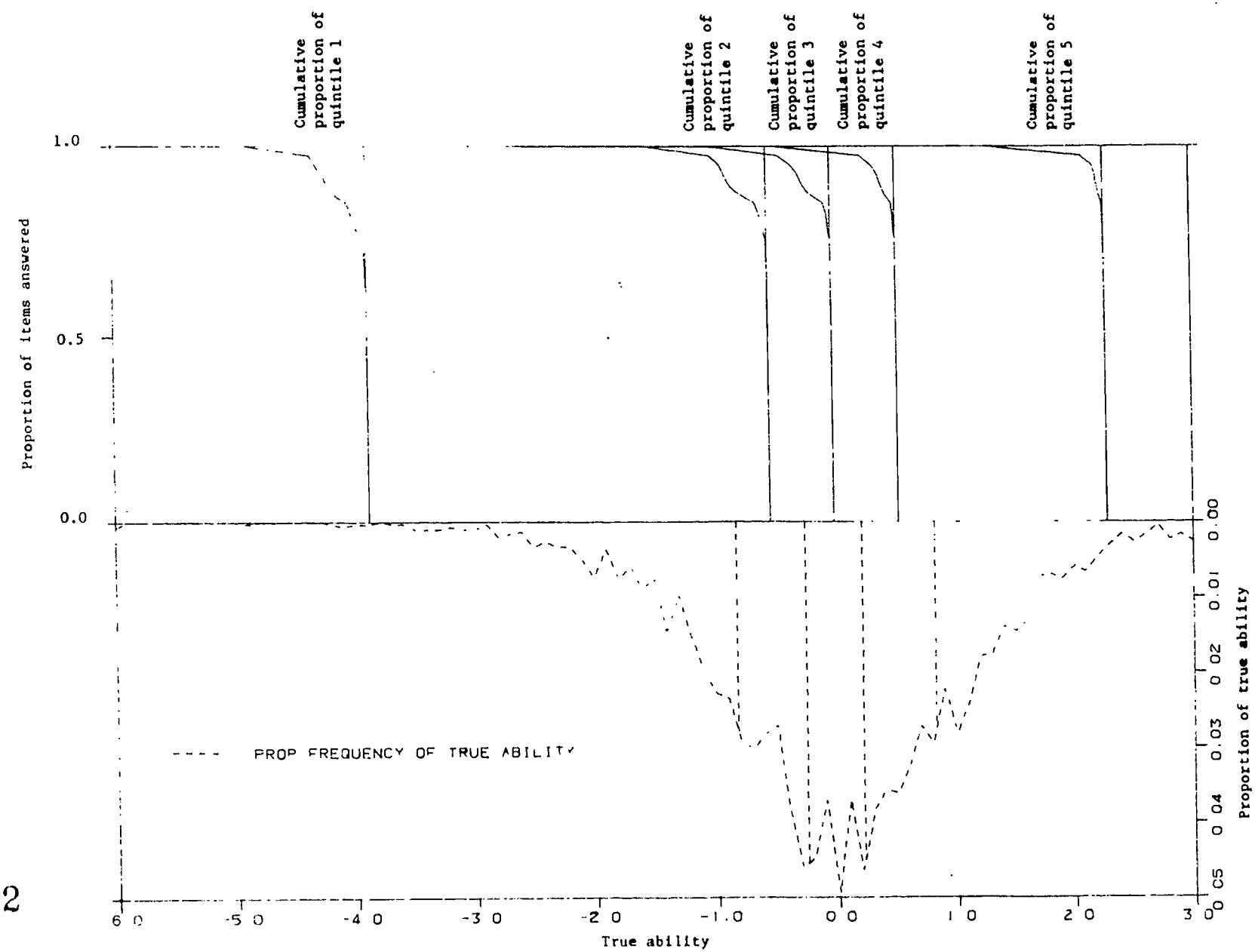


Figure 2. The cumulative distributions of the percentage of items reached by quintile of true ability for SAT-V, Section 1.

Figure 3. The cumulative distributions of the percentage of items reached by quintile of true ability for SAT-V, Section 2.



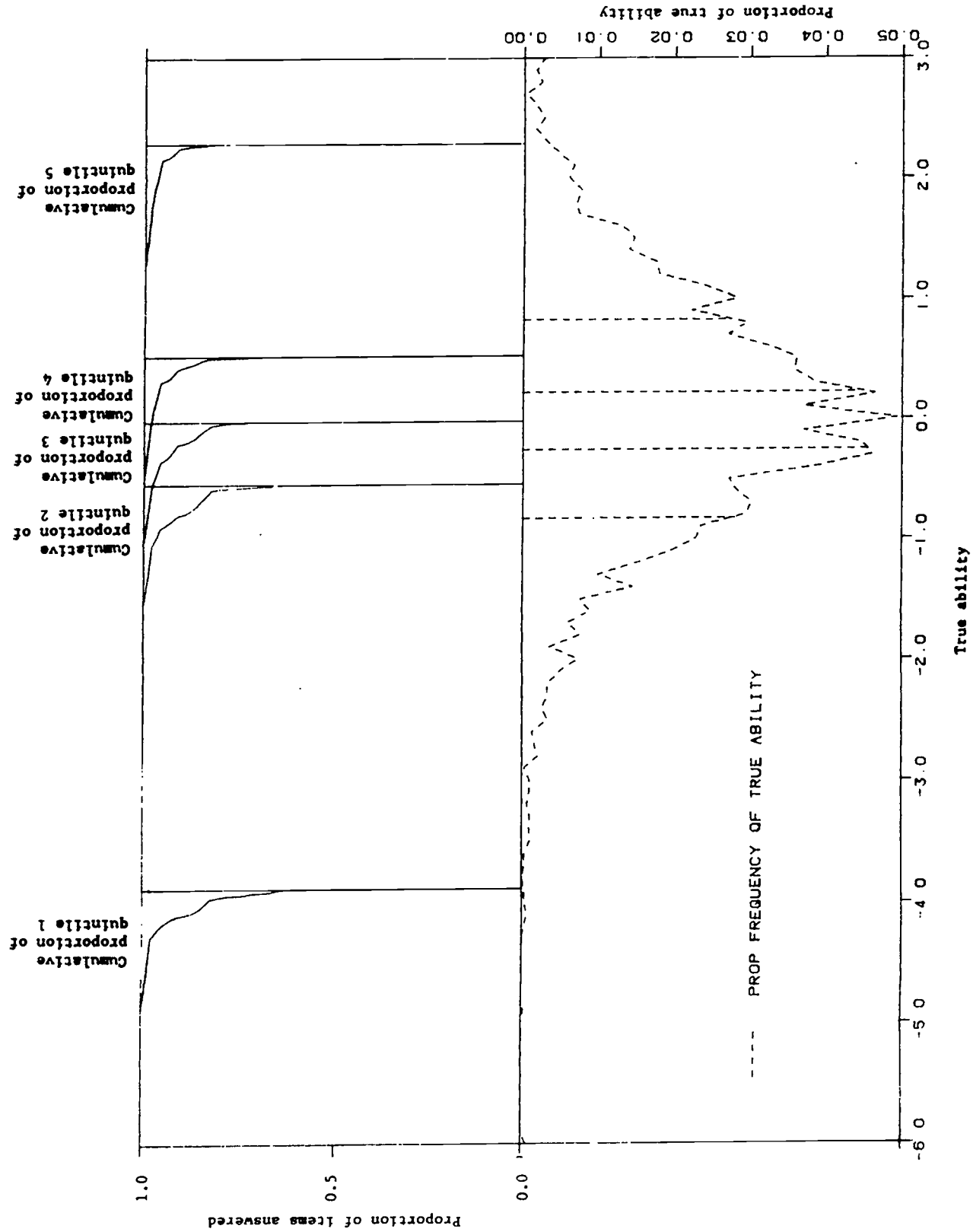


Figure 4. The cumulative distributions of the percentage of items reached by quintile of true ability for SAT-V, anchor test section.

Figure 5. For selected items, the proportion of omits in true response strings, separately by quintiles for right/wrong modeled responses.

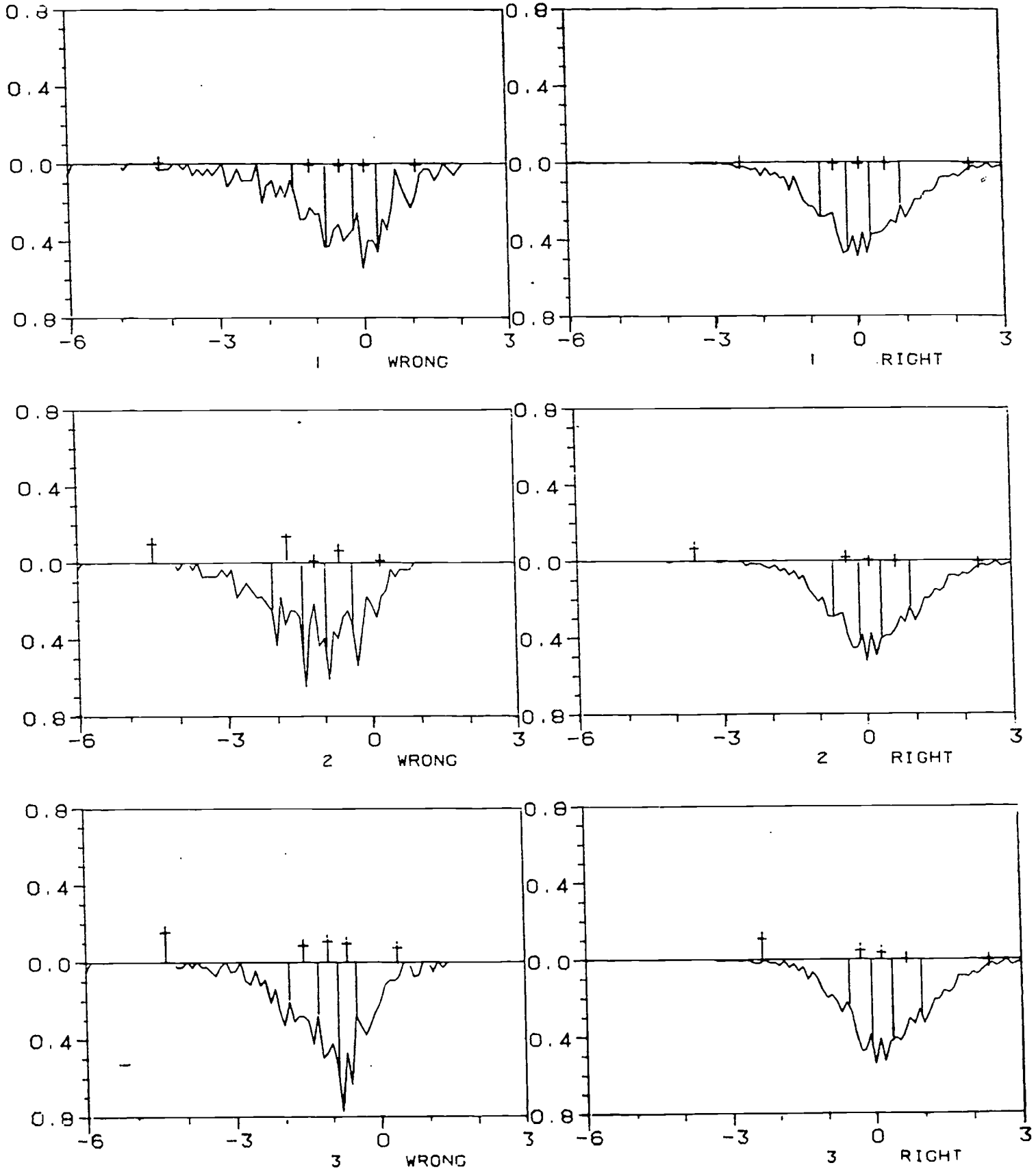


Figure 5, continued. For selected items, the proportion of omits in true response strings, separately by quintiles for right/wrong modeled responses.

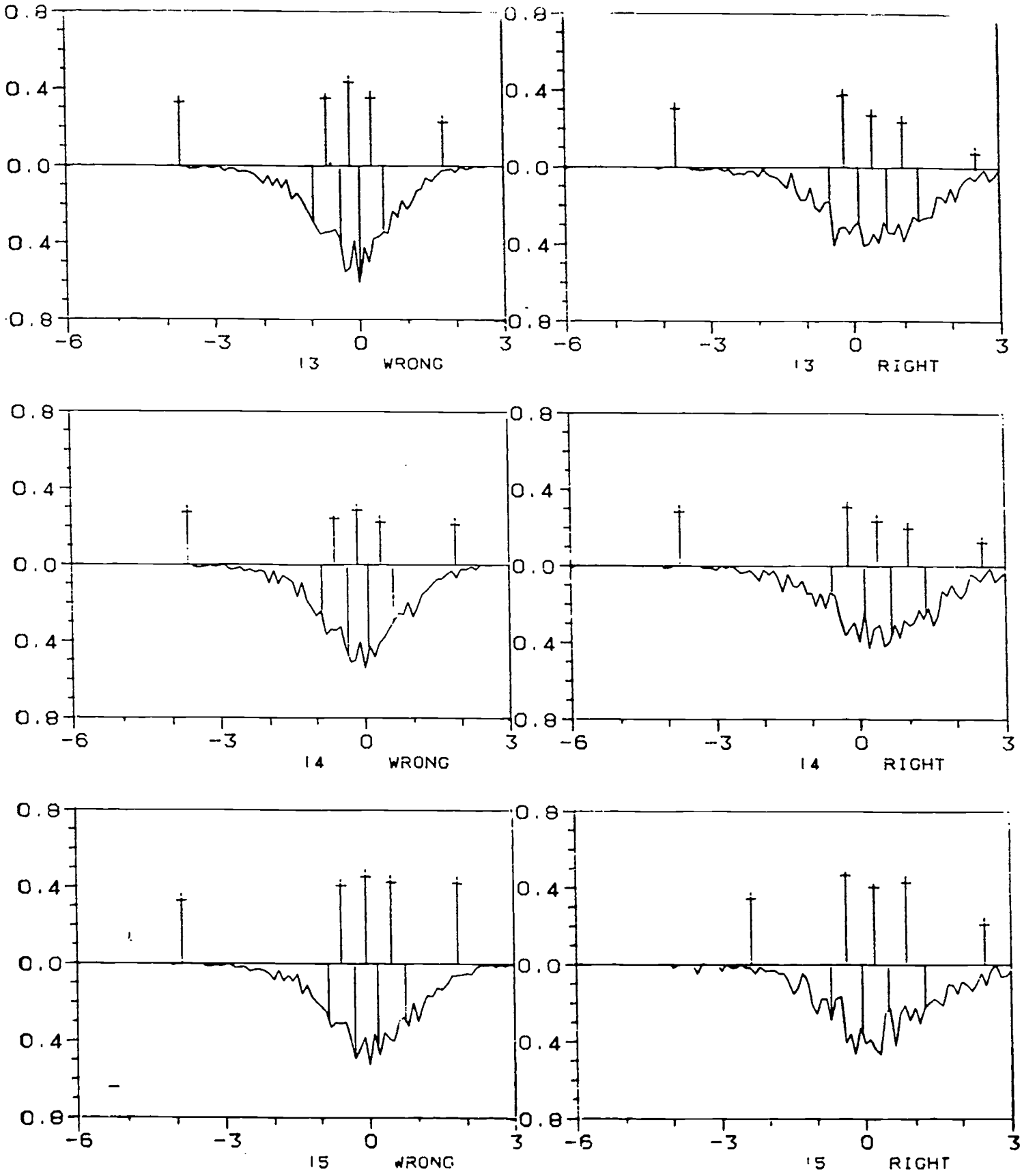


Figure 5, continued. For selected items, the proportion of omits in true response strings, separately by quintiles for right/wrong modeled responses.

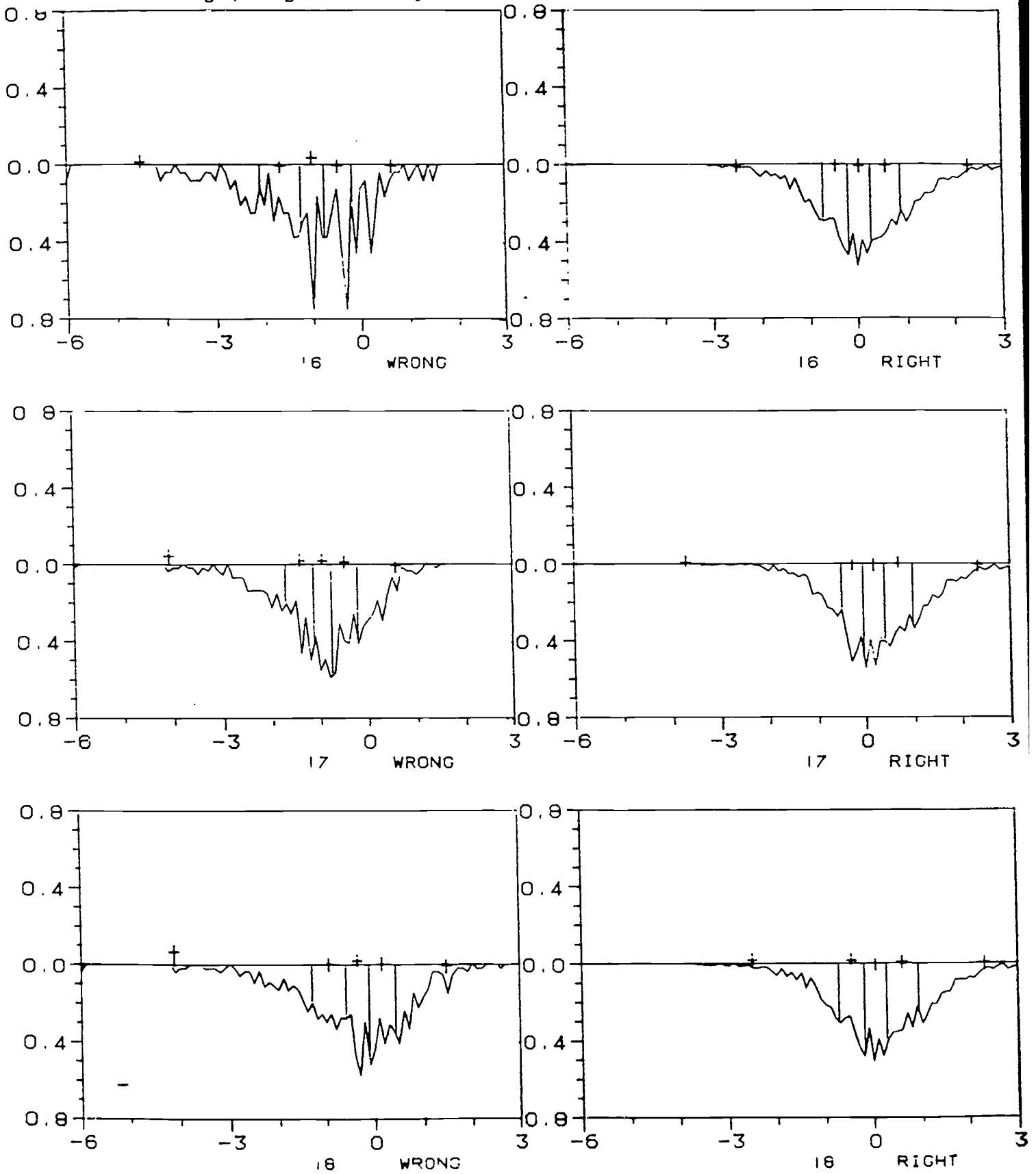


Figure 5, continued. For selected items, the proportion of omits in true response strings, separately by quintiles for right/wrong modeled responses.

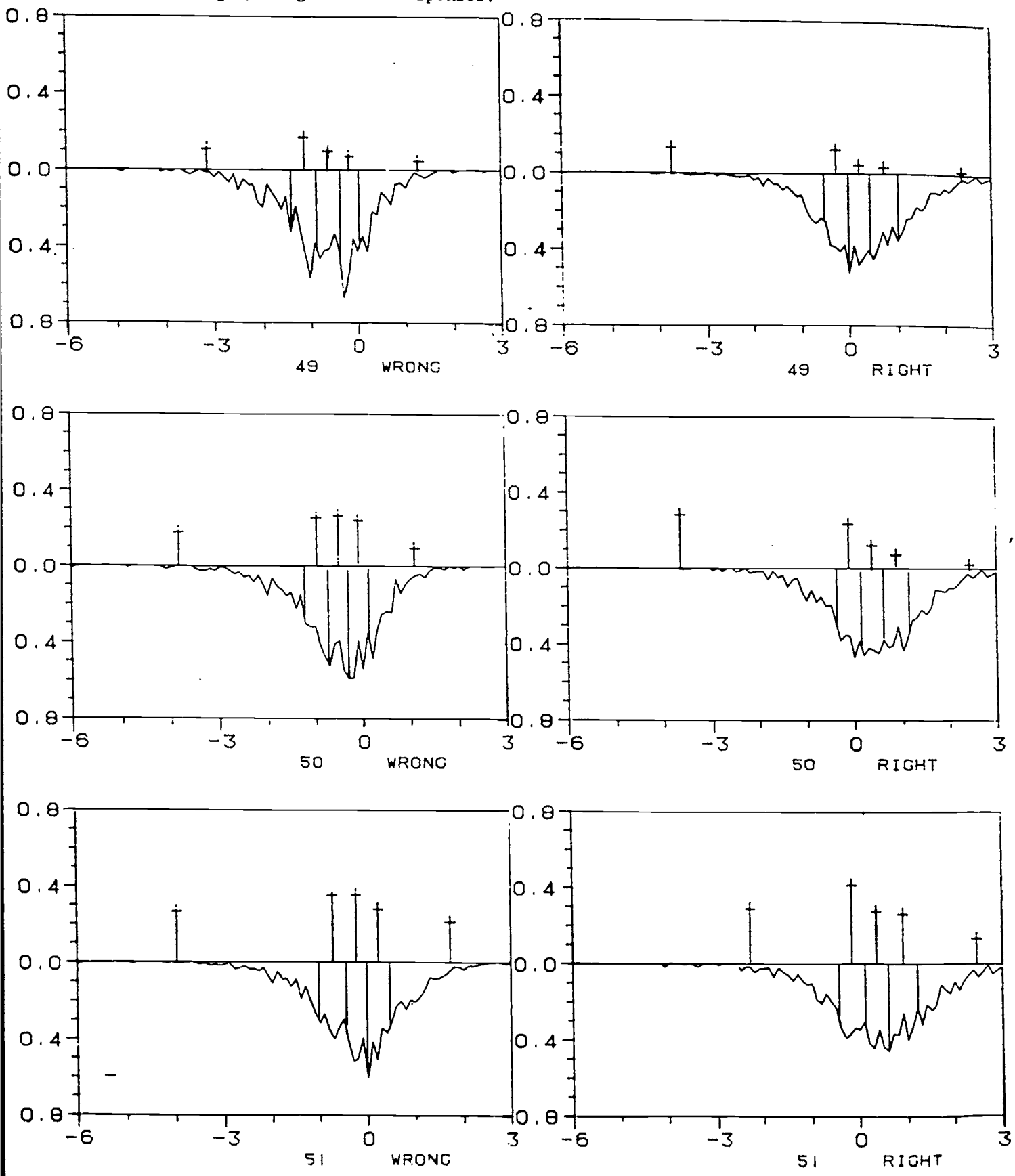


Figure 5, continued. For selected items, the proportion of omits in true response strings, separately by quintiles for right/wrong modeled responses.

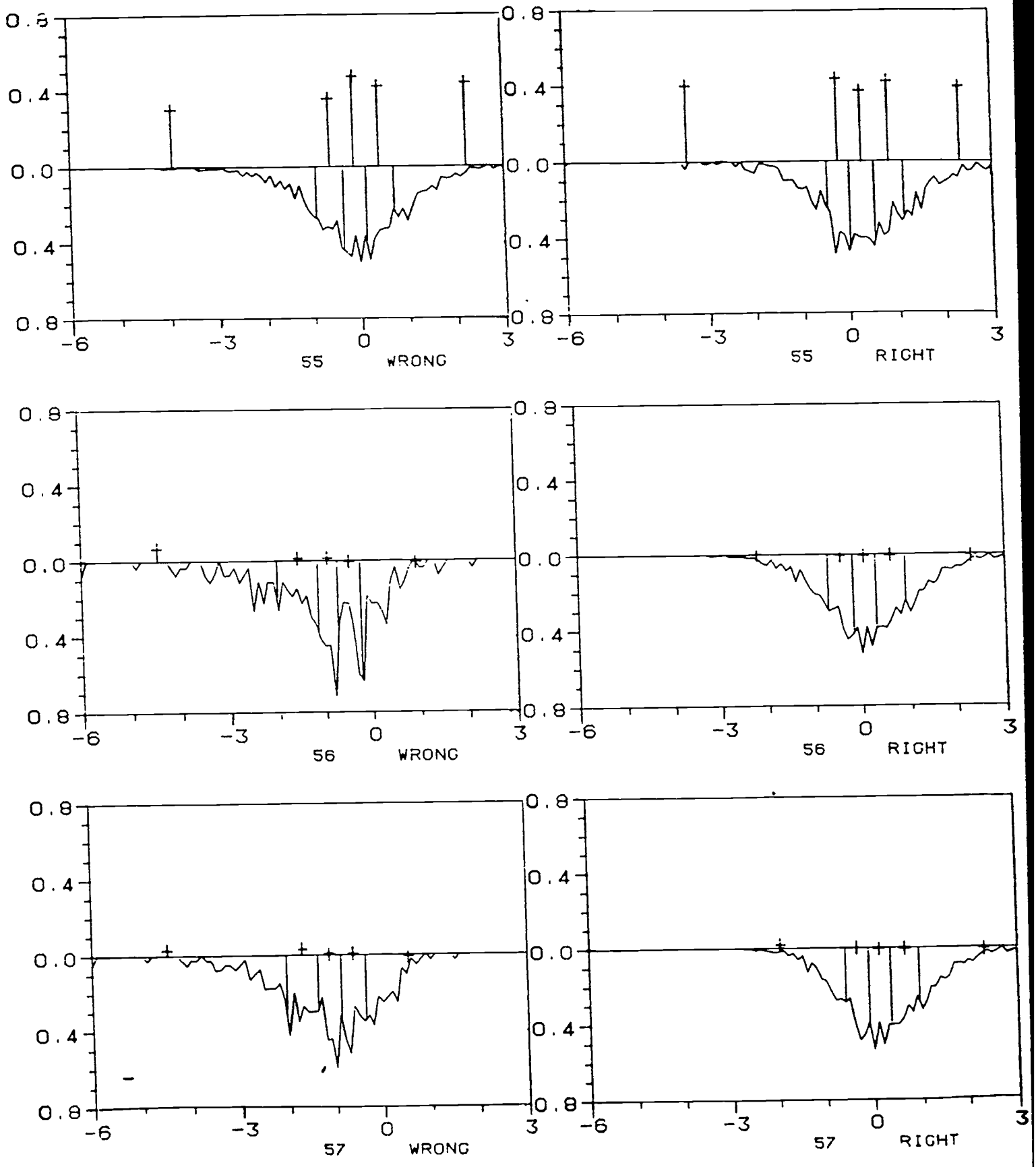


Figure 5, continued. For selected items, the proportion of omits in true response strings, separately by quintiles for right/wrong modeled responses.

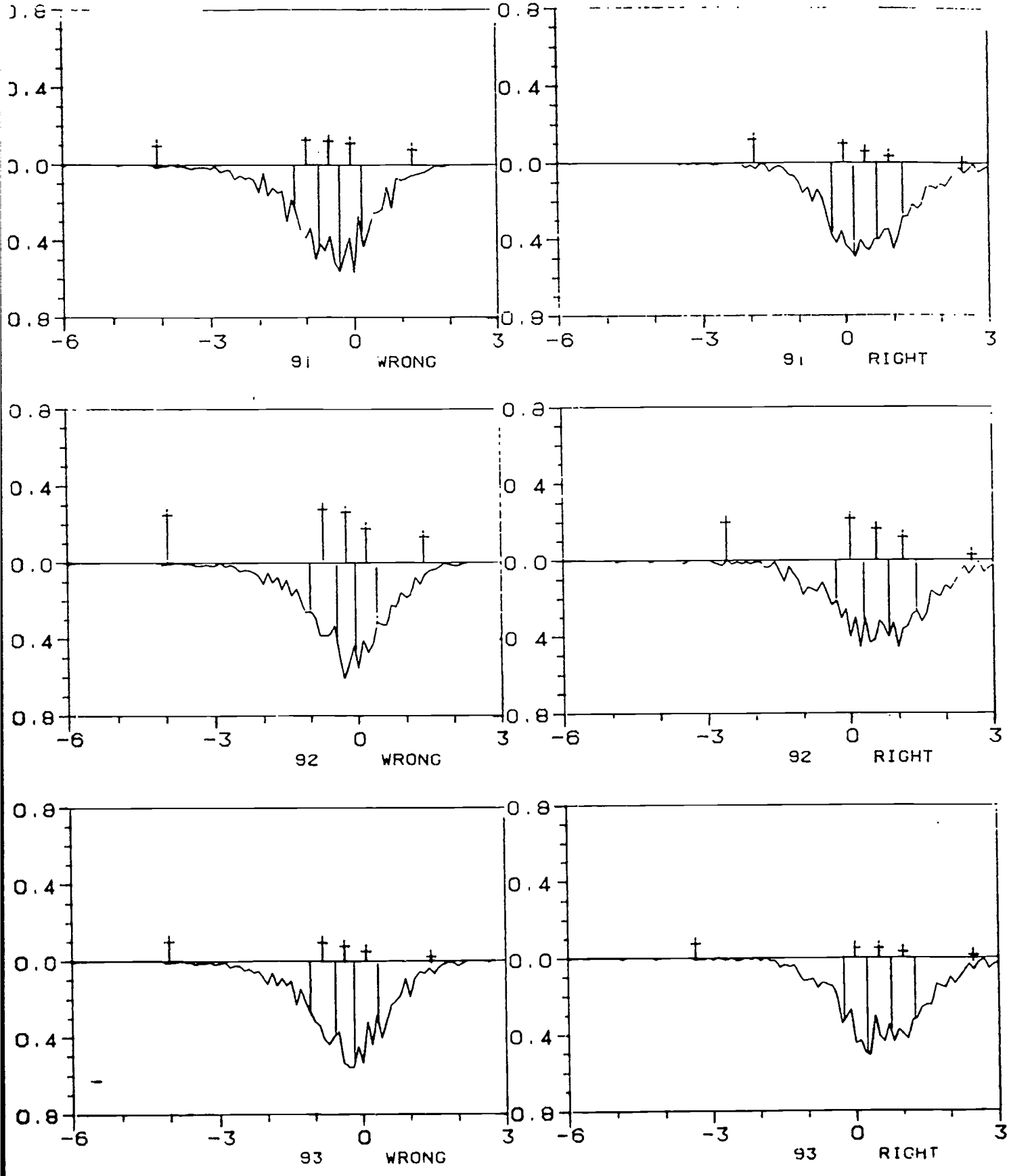


Figure 5, continued. For selected items, the proportion of omits in true response strings, separately by quintiles for right/wrong modeled responses.

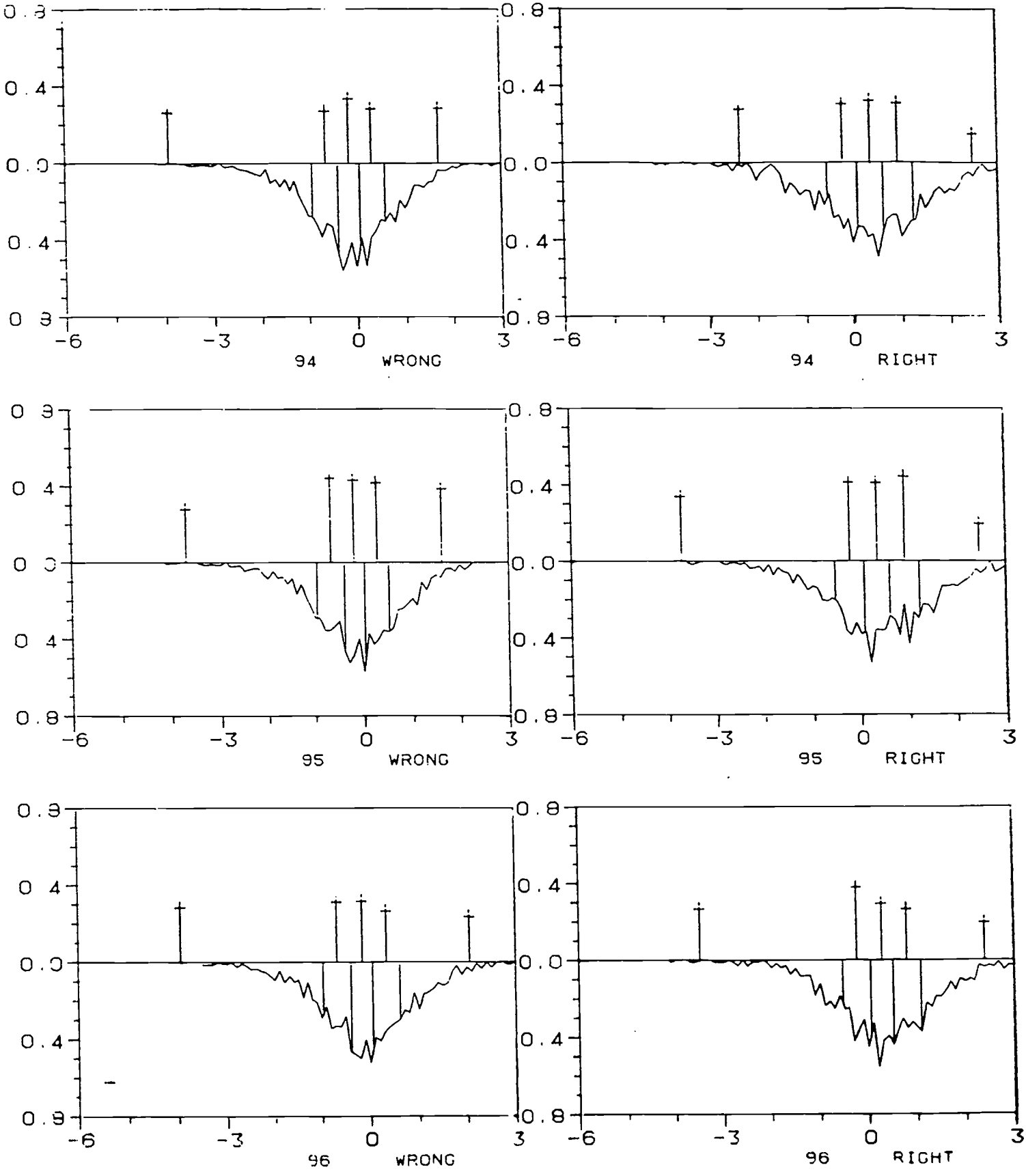
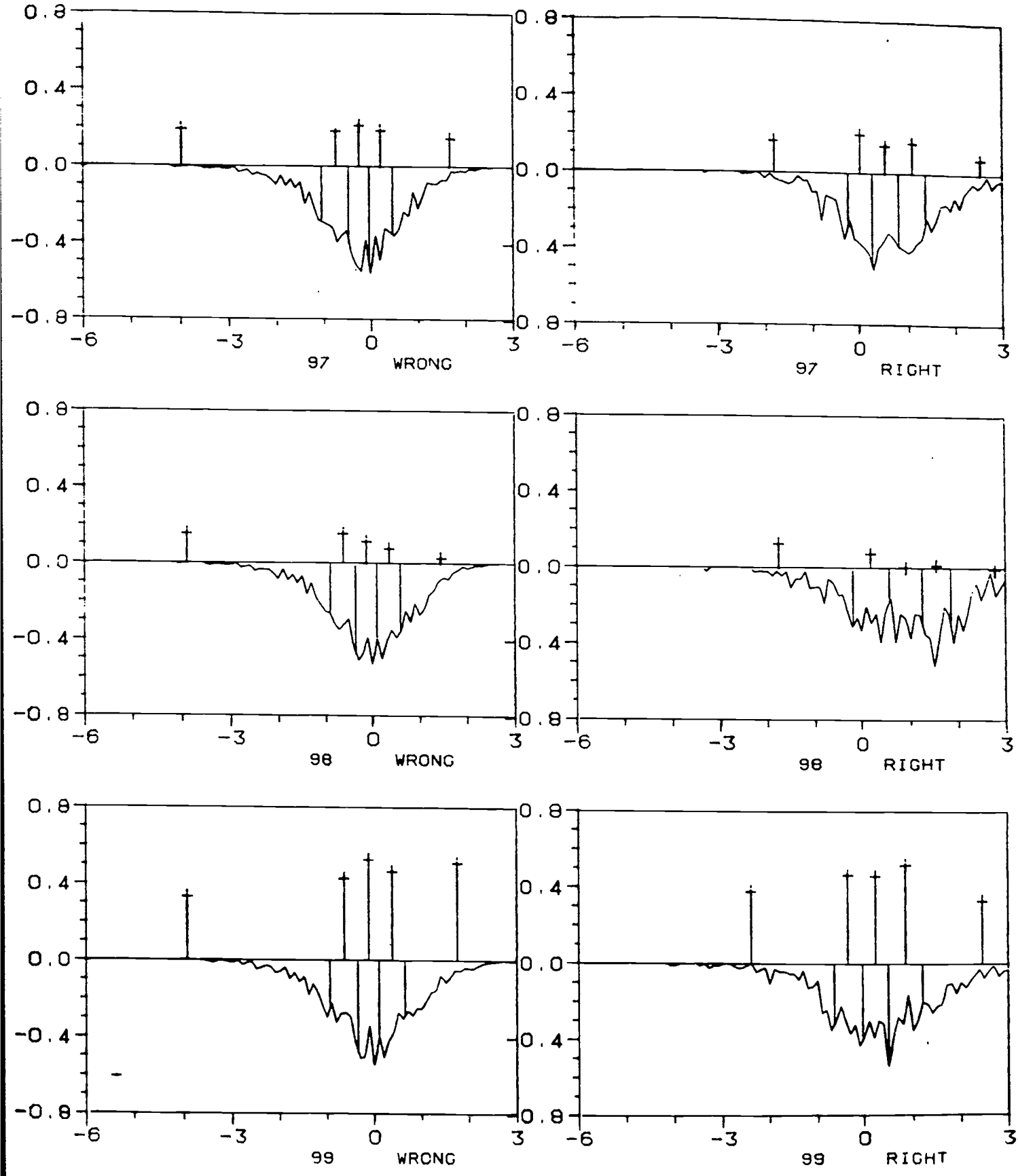


Figure 5, continued. For selected items, the proportion of omits in true response strings, separately by quintiles for right/wrong modeled responses.



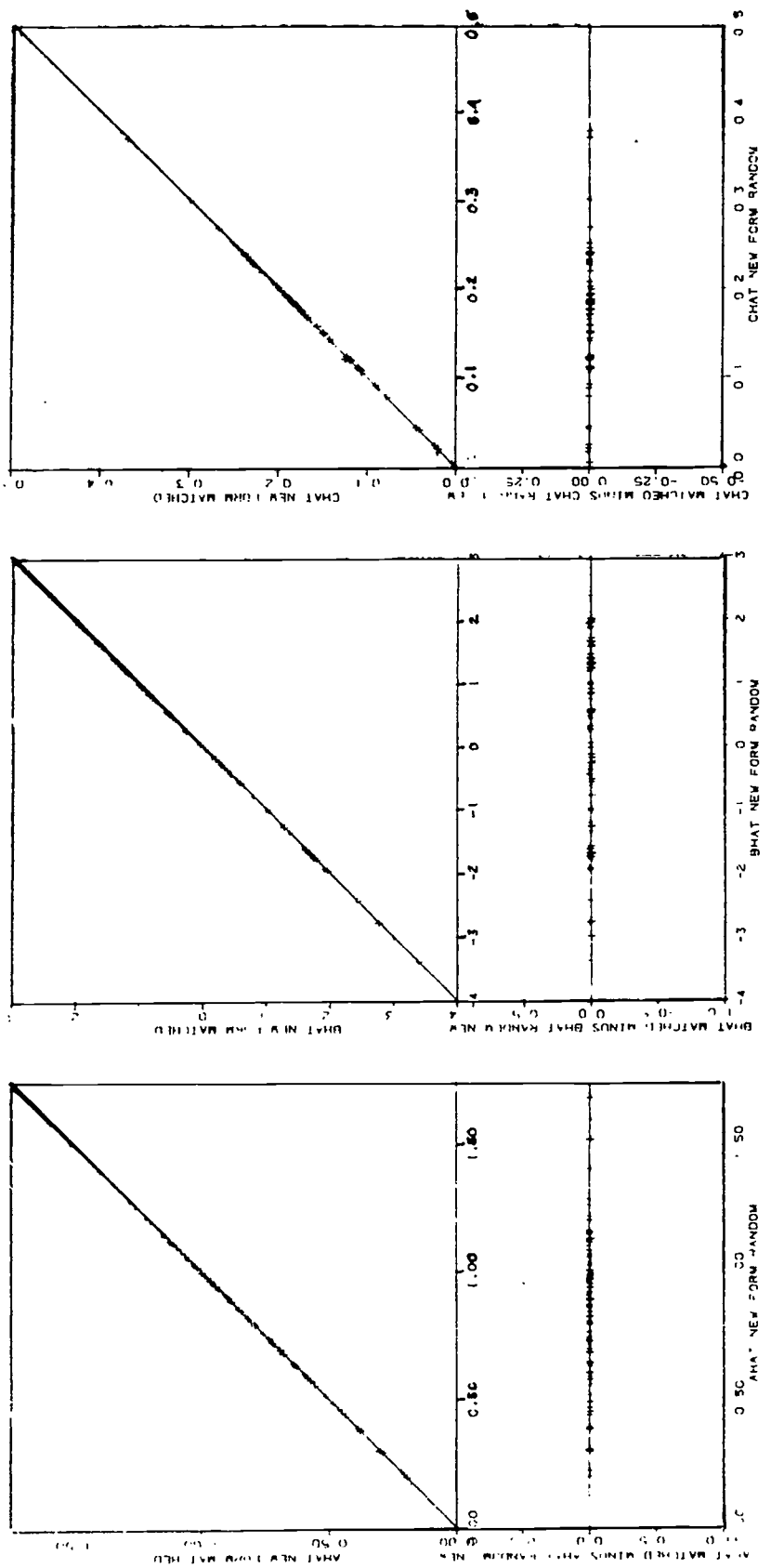


Figure 6a. For simulated data, item parameter estimates for matched vs. random-and-unequal conditions and residuals, for NEW.

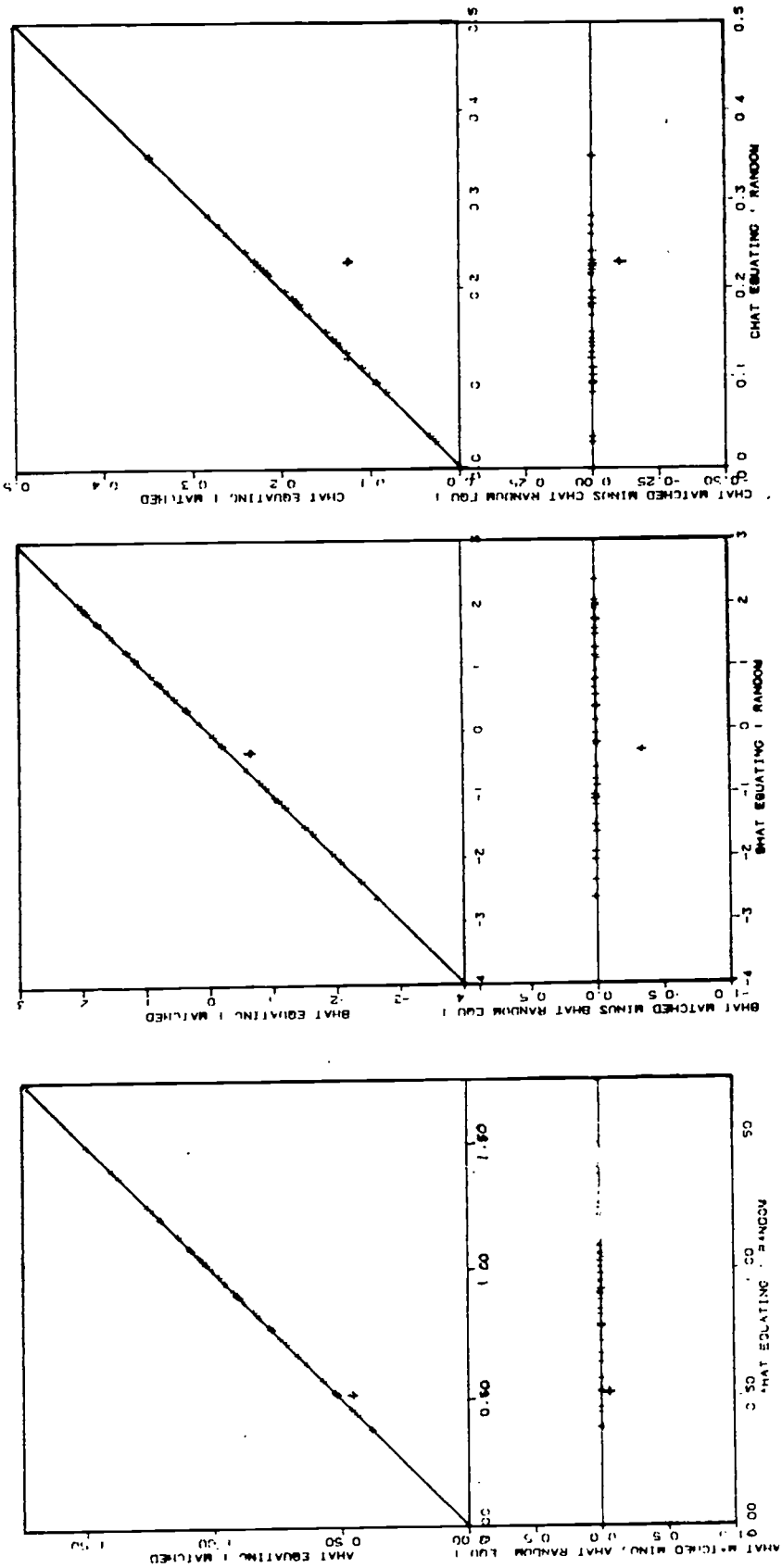


Figure 6b. For simulated data, item parameter estimates for matched vs. random-and-unequal conditions and residuals, for EQ1.

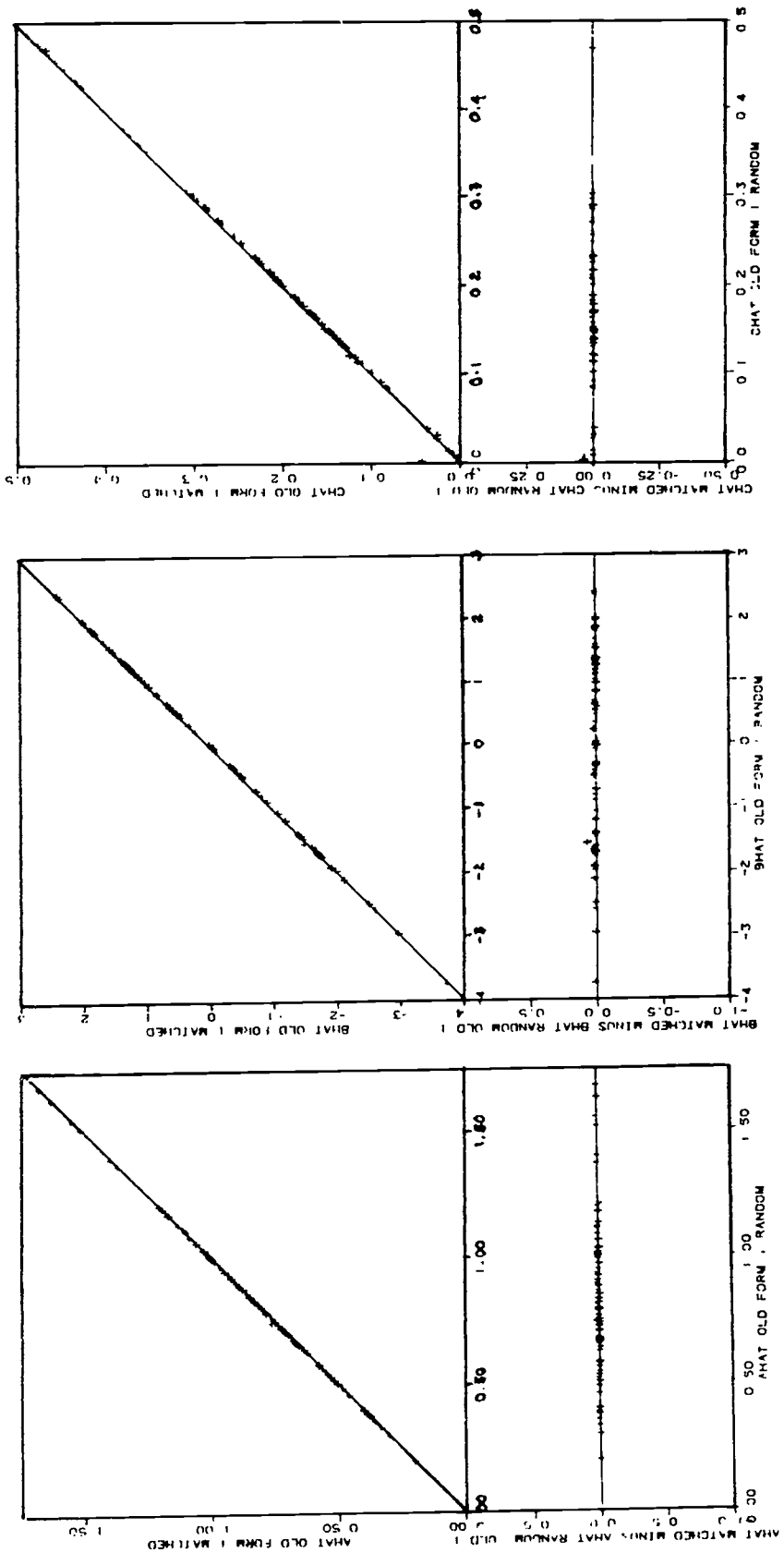


Figure 6c. For simulated data, item parameter estimates for matched vs. random-and-unequal conditions and residuals, for OLD1.

Figure 6d. For simulated data, item parameter estimates for matched vs. random-and-unequal conditions and residuals, for EQ2.

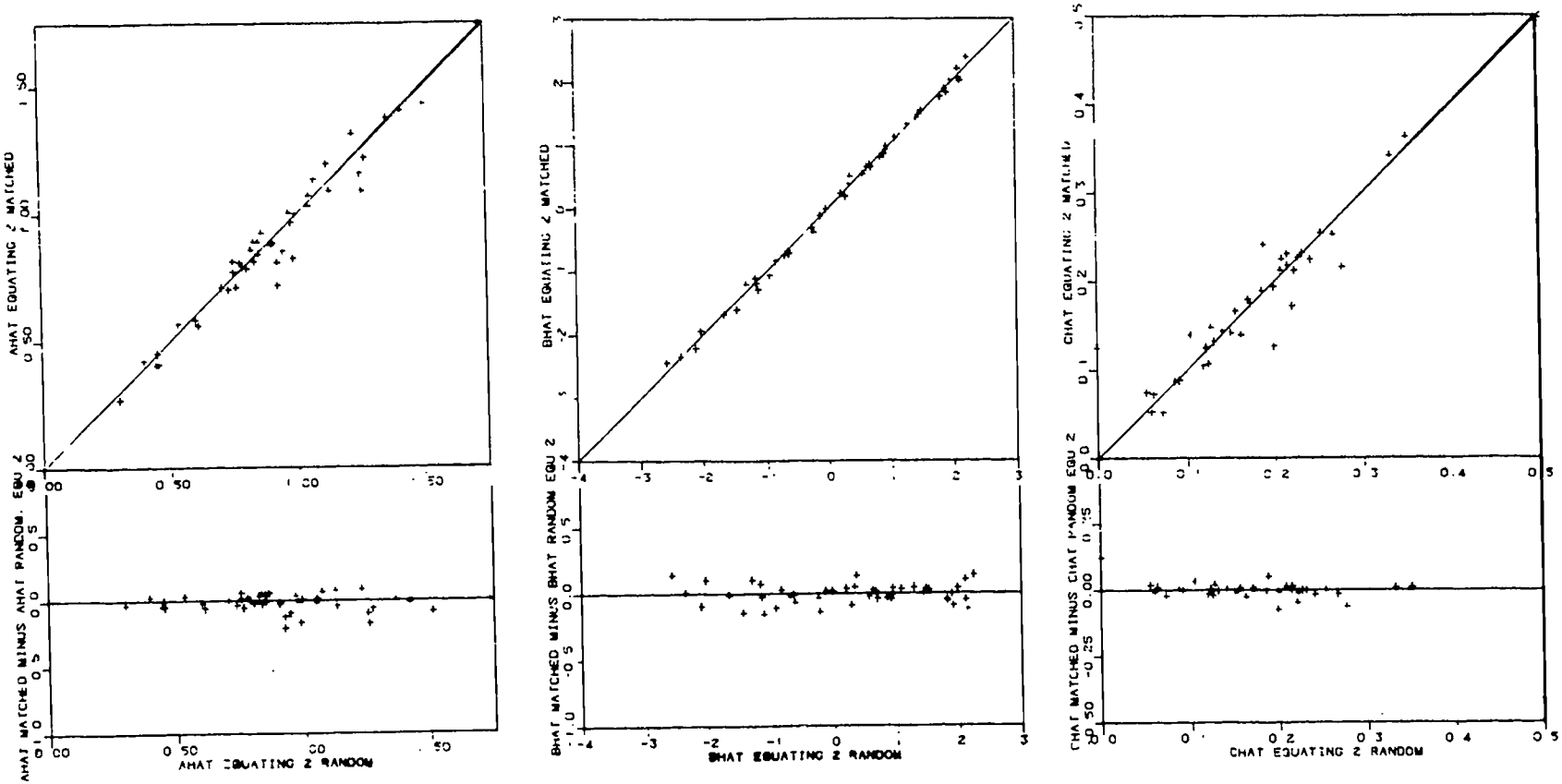
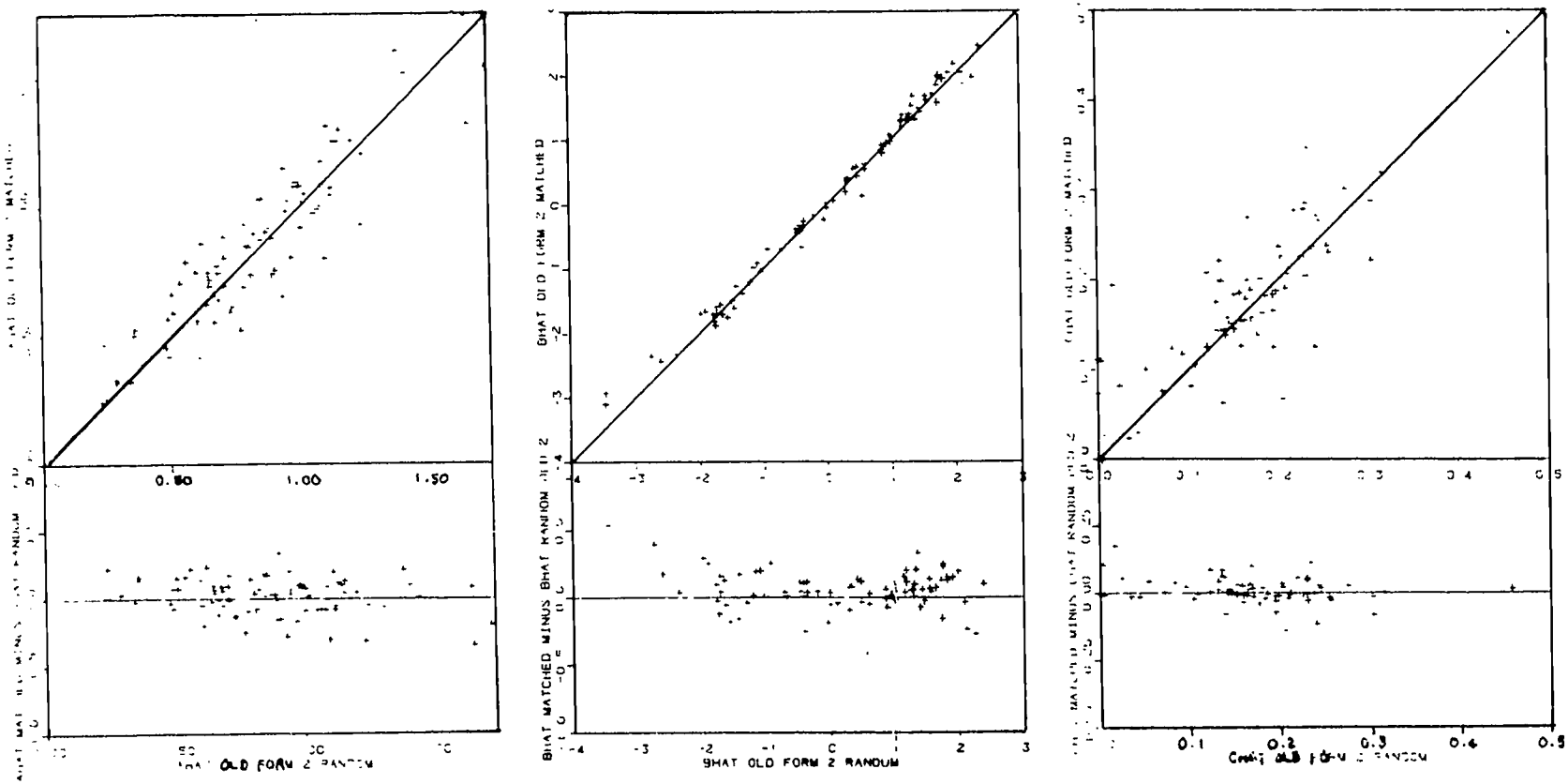


Figure 6e. For simulated data, item parameter estimates for matched vs. random-and-unequal conditions and residuals, for OLD2.



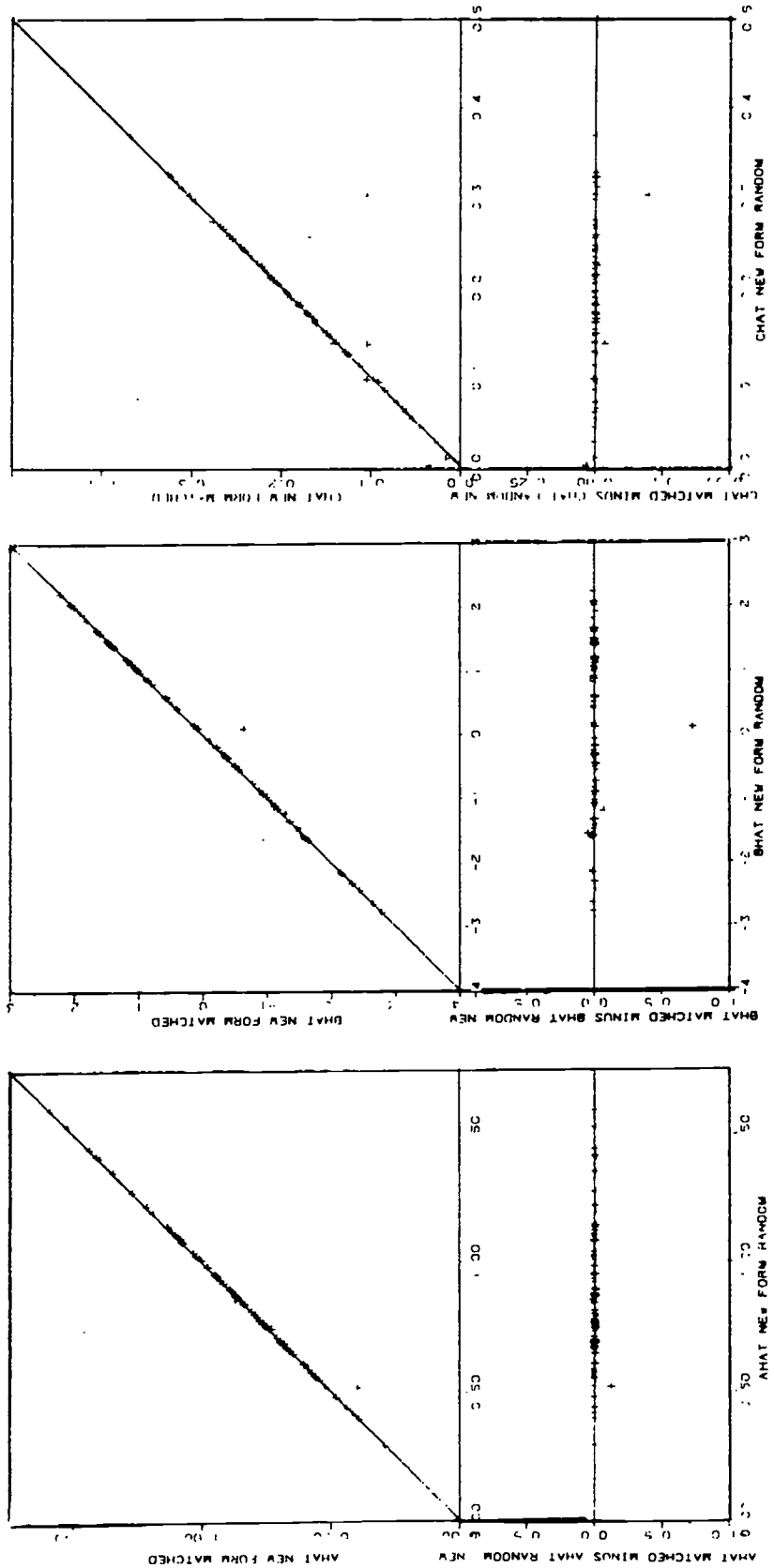


Figure 7a. For real data, item parameter estimates for matched vs. random-and-unequal conditions and residuals, for NEW.

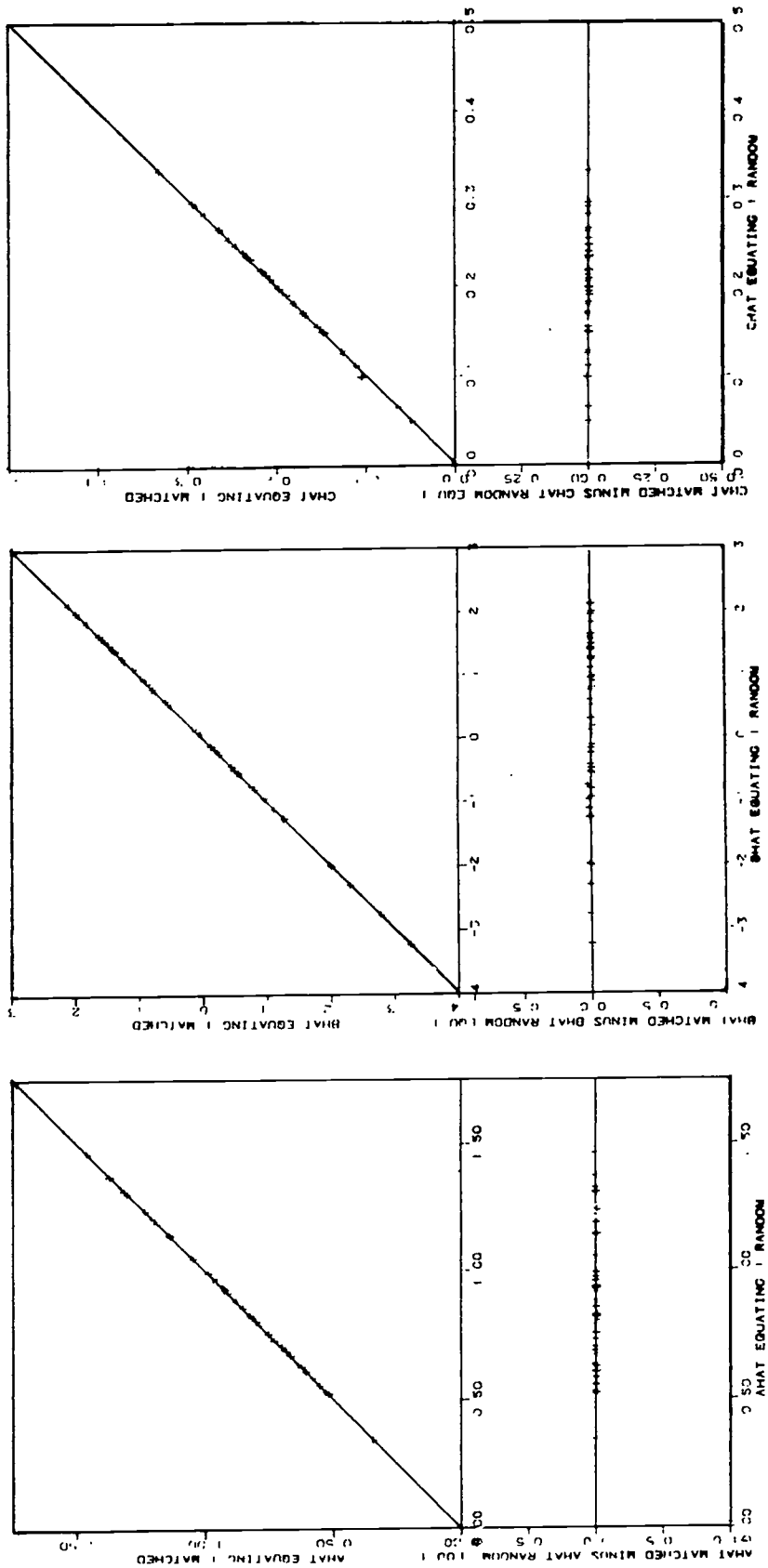


Figure 7b. For real data, item parameter estimates for matched vs. random-and-unequal conditions and residuals, for EQ1.

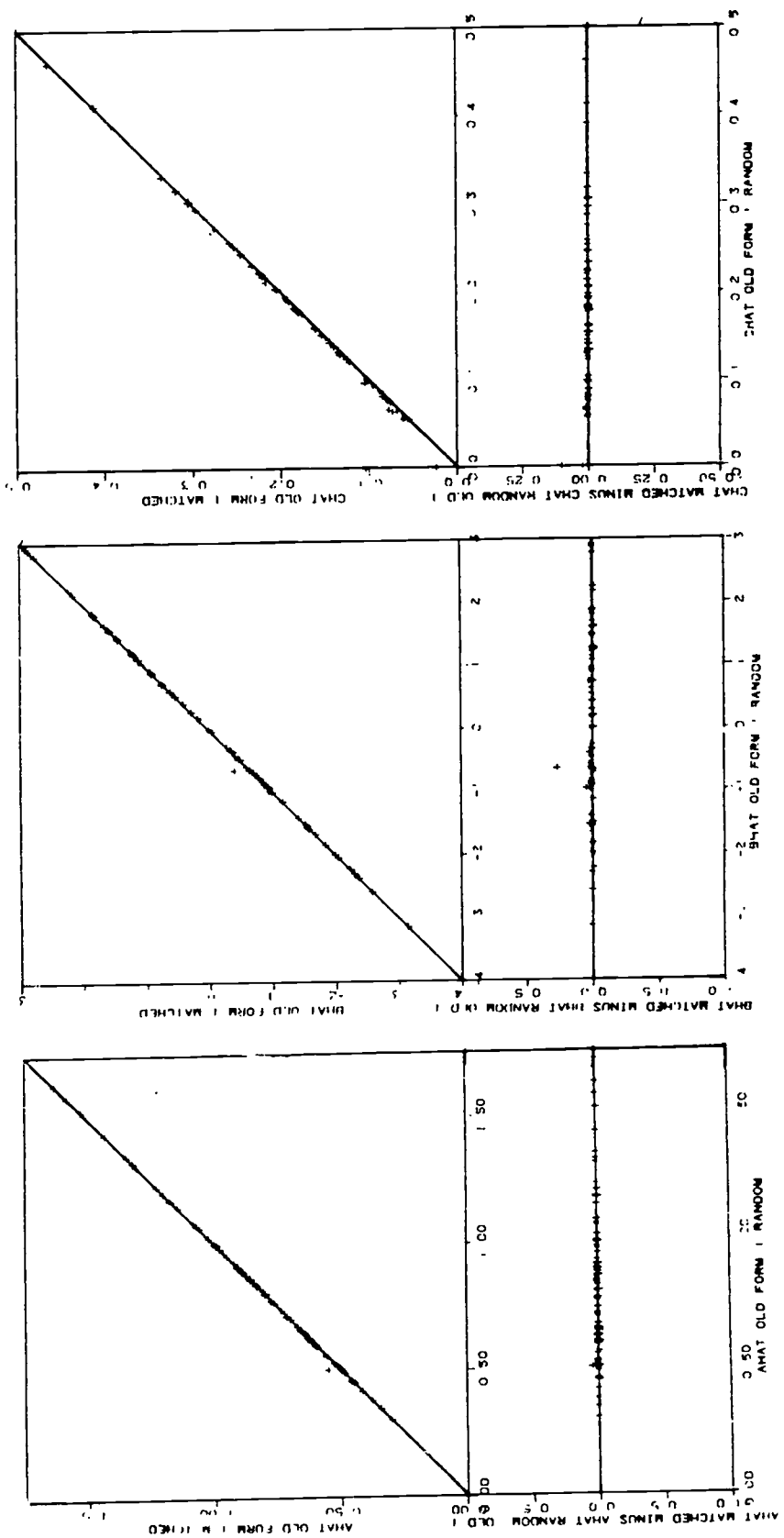
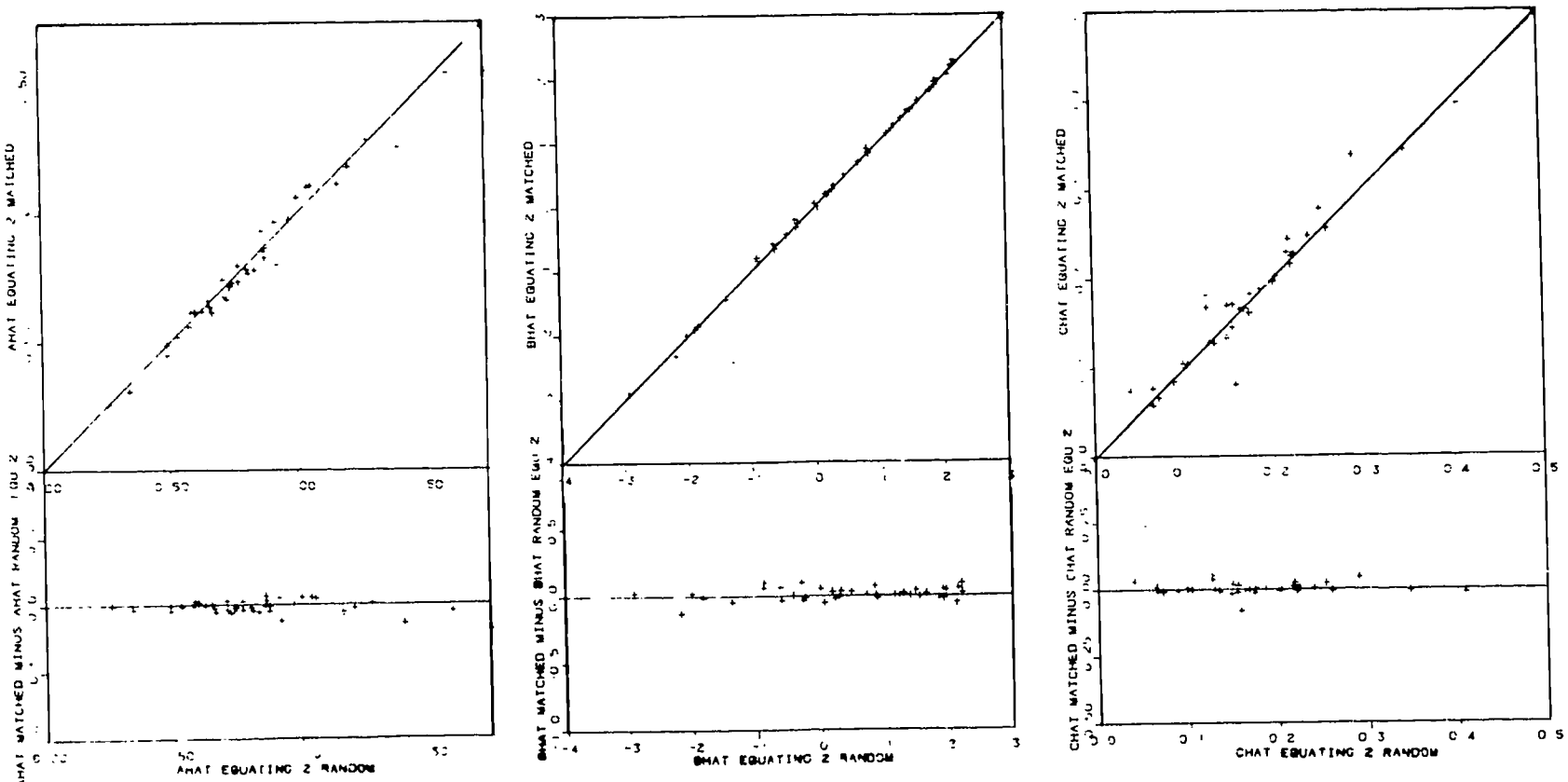


Figure 7c. For real data, item parameter estimates for matched vs. random-and-unequal conditions and residuals, for OLD1.

Figure 7d. For real data, item parameter estimates for matched vs. random-and-unequal conditions and residuals, for EQ2.



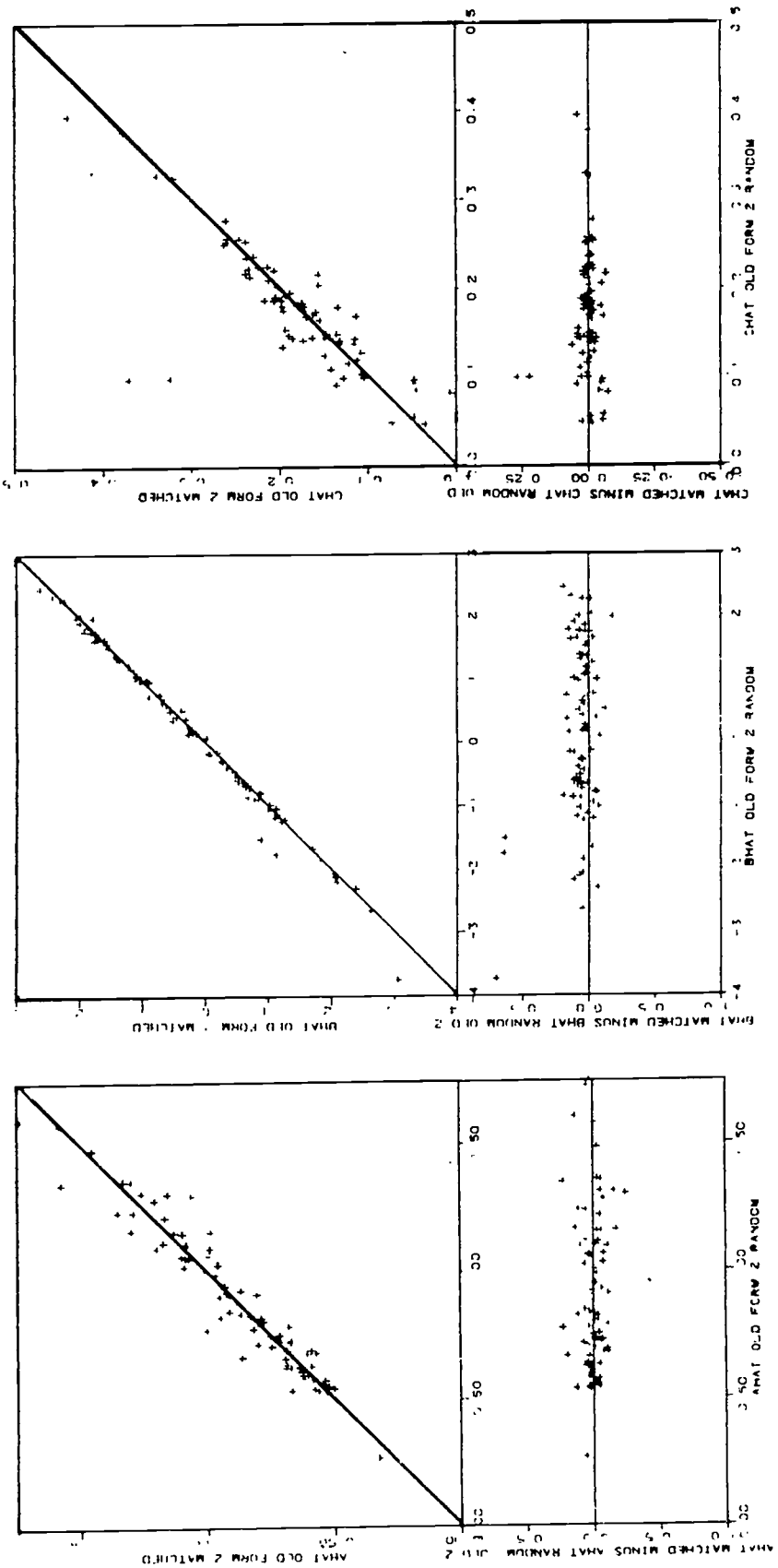


Figure 7e. For real data, item parameter estimates for matched vs. random-and-unequal conditions and residuals, for OLD2.

Figure 8. Projected scaled score means for all equatings methods and all experimantal conditions.